ARMY RESEARCH LABORATORY

Using AutoMap for Social and Textual Network Analysis

by William Tanenbaum and John Brand

ARL-TN-321 **July 2008**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5067

ARL-TN-321 **July 2008**

Using AutoMap for Social and Textual Network Analysis

William Tanenbaum
Science and Engineering Apprentice Program (SEAP)

John Brand
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
July 2008	Final	October 2006–September 2007

4. TITLE AND SUBTITLE
Using AutoMap for Social and Textual Network Analysis

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
William Tanenbaum* and John Brand

5d. PROJECT NUMBER
8TEPRC-SNA

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: AMSRD-ARL-CI-CT
Aberdeen Proving Ground, MD 21005-5067

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TN-321

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
* Mr. Tanenbaum is a student at the Science and Engineering Apprentice Program (SEAP).

14. ABSTRACT
This guide for AutoMap has been developed to allow researchers on the Social Network Analysis Team to rapidly gain facility when using this complicated, subtle, and powerful software package; to avoid some of the same common problems and pitfalls; and to gain some measure of common method. It is a resource to be used in conjunction with the various users' guides provided with the AutoMap software. This guide has been developed using AutoMap version 2.6.70. It will be revised as AutoMap evolves and as the team gains expertise.

15. SUBJECT TERMS
AutoMap, social networking, data mining, textual analysis

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. John Brand
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL	44	19b. TELEPHONE NUMBER (*Include area code*) 410-278-4454

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

Acknowledgments

This guide is the result of pooling information from members of the Social Network Analysis (SNA) Team.

The SNA Team is composed of Janet O'May, team leader, John Brand, Ashley Foots, Joan Forester, Sean Murray, and William Tanenbaum.

1. Background

Social and textual network analysis is an important topic in a number of disciplines that may be used to provide information support to combat units. One tool used for performing social and textual network analysis is a free package of software developed by the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University.[1] One tool in the package of software is AutoMap.[2]

The use of AutoMap in students' research is particularly difficult due to the sporadic nature of the students' involvement and their constant entry into and departure from a given lab or project within a lab. Additionally, if each researcher has to learn, usually the hard way, how to use this package by running it over and over, an enormous amount of time is lost and the opportunity for mistakes grows rapidly. Simply, there is not sufficient time or continuity in practical terms to reinvent this particular wheel over and over again.

This user's guide has been developed to allow researchers to rapidly gain some facility when using this complicated, subtle, and powerful package; to avoid some of the same common problems and pitfalls; and to gain some measure of commonality while using this package in the research under way in this laboratory. It should be considered as a resource to be used in conjunction with the various user's guides provided with the software.

AutoMap has a complicated and extensive user's guide that outlines how to run AutoMap in great detail.[3] This extensive hyperlinked document is invaluable, but it does not show how to use AutoMap in specific research applications. This learning process takes considerable time, and every analyst will develop somewhat different ideas concerning use, even for the same project. These differences are invaluable, but the time constraints for total learning independence are intolerable in practice.

This guide on using AutoMap has been developed using AutoMap version 2.6.70. It will be revised as AutoMap evolves and as the Social Network Analysis (SNA) team gains expertise in using AutoMap in the team's particular research projects.

[1] CASOS. http://www.casos.cs.cmu.edu/ (accessed 29 August 2007).

[2] CASOS. http://www.casos.cs.cmu.edu/projects/AutoMap/ (accessed 29 August 2007).

[3] CASOS. http://www.casos.cs.cmu.edu/projects/AutoMap/software/2.6.60/help/AutoMap_2.0 _users_guide.html (accessed 29 August 2007).

2. Text Processing

AutoMap is a program that determines relationship between concepts within a text by analysis of the text strings proximity representing those concepts. The text strings may be single words or groups of characters not separated by spaces. A concept may represent an abstraction or concrete thing—person, place, action, object, number—or simply perform a grammatical, stylistic, or other linking function. Figure 1 illustrates the way in which this process is done. This figure sketches the user's actions in transformation of a source text as it is prepared for analysis ("preprocessed") and then analyzed.

It is necessary to "preprocess" the textual material to eliminate unwanted or unneeded concepts before analysis. This preprocessing involves deletion of unneeded concepts that obscure meaning and transformation of multiple word or symbol groupings that refer to a single idea into unbroken text strings. There are other preprocessing utilities that are discussed in this report, but the most important preprocessing functions involve the Delete List and the Generalization Thesaurus. There are other thesauri that the user may use to assign meaning or associate concepts; those are discussed in this report but not shown in figure 1.

Eliminating unnecessary material with Delete Lists and transforming multiple word groupings into unbroken text strings with the Generalization Thesaurus depend on the purpose for which the textual analysis is performed. A concept central to an analysis to explore one set of material may be irrelevant to another set. For this reason, different analyses or different types of source material may require specialized and unique delete lists and thesauri. When the input text has been sufficiently prepared, the concepts are associated as "bigrams," or pairs of character strings. The association is based on proximity—concepts located next to each other in the processed text are associated in pairs as bigrams. The meaning behind the association must be determined separately.

AutoMap has several preloaded thesauri. The analyst may configure a thesaurus specific to the analysis starting from a listing based on the source document text at various stages of preprocessing, use a previously developed thesaurus, or define a new thesaurus. It may be mentioned in passing that one of the pitfalls of using several thesauri, which is not mentioned in the user's guide, is that each time the analyst shifts to a new thesaurus the program automatically undoes the application of the previous thesaurus. Thus if one desires to apply several special-purpose thesauri, the material in the several thesauri must be merged off-line. This is a good example of a pitfall that each user must discover separately, unless cautioned by a document such as this.

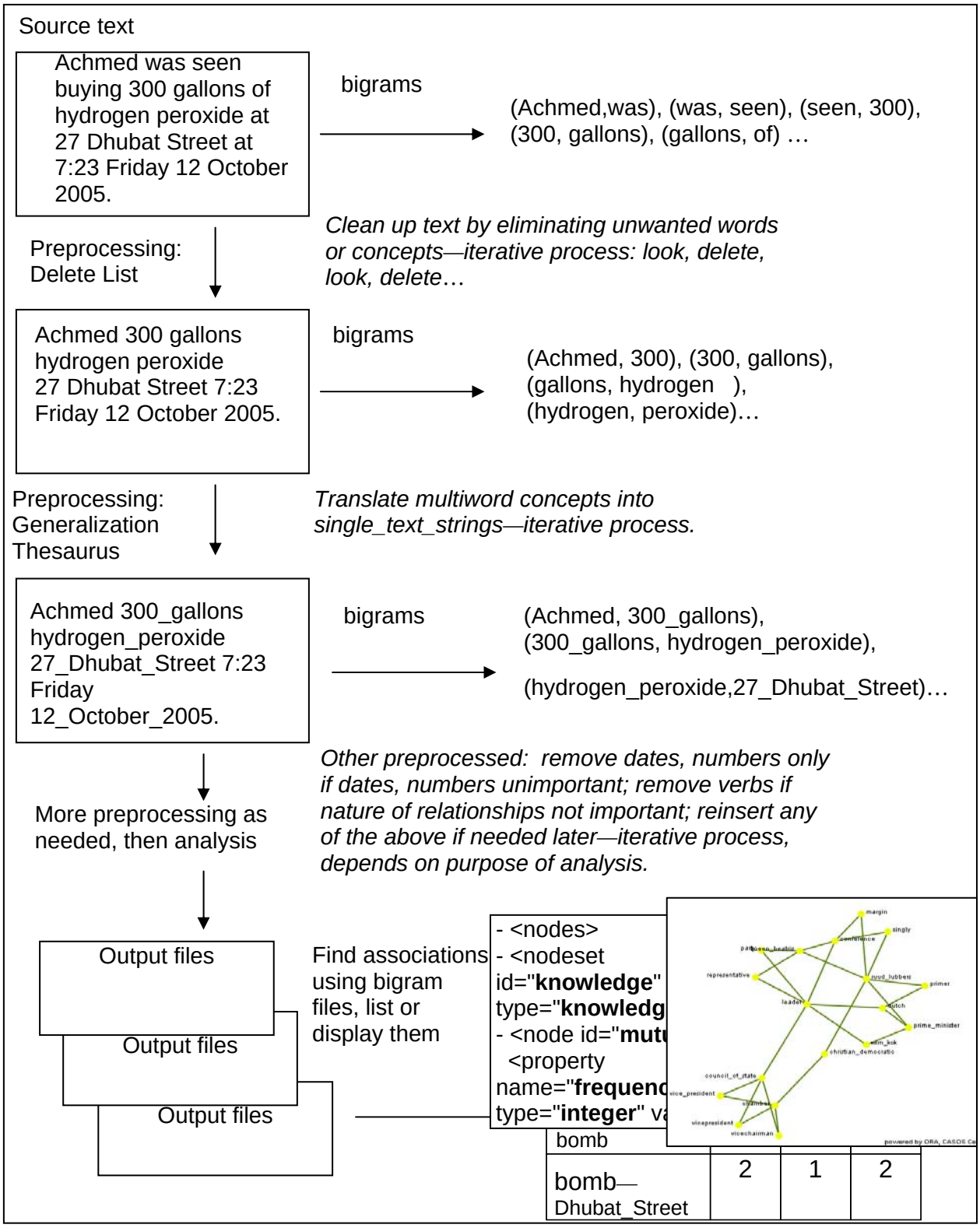


Figure 1. Schematic outline of how AutoMap functions.

Other thesauri allow the user to specify meaning in terms of concept function, both individually and generally. Specific function may be assigned to individual concepts by use of a Meta-Matrix Thesaurus. The Meta-Matrix Thesaurus allows the user to assign a function to a concept. That is, "Dhubat Street" is assigned to <location>, and "Achmed" to <agent>. This is particularly useful in constructing an ontology based on AutoMap output. The user may also define functions, in addition to using the predefined ones.

An additional thesaurus is the Sub-Matrix Thesaurus, a way of grouping relationship sub-networks by meta-matrix categories.

A relationship between concepts, as revealed by AutoMap, is based on proximity of the strings representing them in the preprocessed text. The exact nature of the relationship (physical position, logical dependence, temporal or sequential, causal, etc.) has to be elicited by reference to a third concept, usually a verb or other part of speech. The verb or other part of speech may be removed inadvertently by inclusion of the concept in the Delete List. This is an obvious pitfall for the analyst. For example, "dog bites man" has a different meaning from "man bites dog," and a Delete List that removes "bites" from the text completely destroys the key factor in the relationship between "man" and "dog."

Another pitfall to understanding meaning is the expression of a trigram as two bigrams to include the central relational term: a text including "man bites dog" and "man bites tomato" will be summarized as two occurrences of the bigram "man bites" and single occurrences of the bigrams "bites dog" and "bites tomato." This averages out key meaning. To reveal that meaning requires some work with the processed text, which militates against successful automation.

Further, if the directionality of the set of bigrams, or pairs of juxtaposed terms, is averaged out, the remaining cue to meaning that is reflected in word order ("man-dog" vs. "dog-man") is lost as well.

Once preprocessing is complete to the user's satisfaction, the textual analysis may be performed.

3. A Note on Using This Guide

This guide leads the analyst through the program application and references what to do and what can go wrong. Menu items are represented by [brackets], and relationships between menu items or sequential application of menu items by ->. Files are represented by use of curly brackets as {file}. Specific text or character strings are represented by quotation marks to distinguish between this narrative text and the subject string; these quotation marks are not used in the program as such. Thus, reference to the term or string Dhubat Street is done in this text as "Dhubat Street," but the string is represented in use in AutoMap as Dhubat Street.

The extensive AutoMap User's Guide supplied with the program is a necessary companion and should be used in conjunction with this guide. The Tools accessed from the AutoMap menu are listed as having their own user's guides, accessed through the Tool [Help] menu. However, the [Help] selections in some of the Tools that can be accessed from AutoMap may be located at unexpected file locations that are not linked to the AutoMap file structure. If the Tool Help menu returns a "file not found" message, search for the file using Windows [Start] -> [Search].

Notes are cautions to avoid an error of some sort. The errors were usually discovered by making them. The notes are listed in boldface type.

The use of the program follows a set of tasks shown in figure 2. These tasks are often iterative in nature. For instance, once a Delete List is applied, the preprocessed text may be checked to see if other material is, in that analysis, extraneous. If it is, the extraneous material may be added to the Delete List and the List applied again. The user may then check again, etc.

4. Using the Program/User Task Sequence

The user has a series of actions or tasks that must be completed. AutoMap requires a great deal of necessary user input, which means that the user must make choices that substantially impact the final result. The action sequence is, to a degree, a matter of informed choice as well. For instance, application of the Generalization Thesaurus after application of the Delete List makes for a simpler set of choices. Simplicity is desirable; however, applying the Thesaurus before the Delete List may result in some transformations that would otherwise not occur. The Generalization Thesaurus may include multiple-word terms, one or more components of which are included in the Delete List. Application of the Delete List first breaks those multiple-word terms up, and the Thesaurus does not find them. Thus the broken multiple-word concepts do not appear in the final analysis results. This may be a serious omission or may not matter at all, depending on the situation.

In this report, the sequence of actions must be considered a guideline that has worked for several analysts; the choice may be modified by other analysts faced with different problems as desired to fit the task. It should be, however, an informed choice. For the purposes of developing this guide, the actions and the order in which they were applied are as follows:

1. Define and apply file-saving choices.
2. Initialize the specific analysis.
3. Define and apply Delete Lists.
4. Define and apply a Generalization Thesaurus.

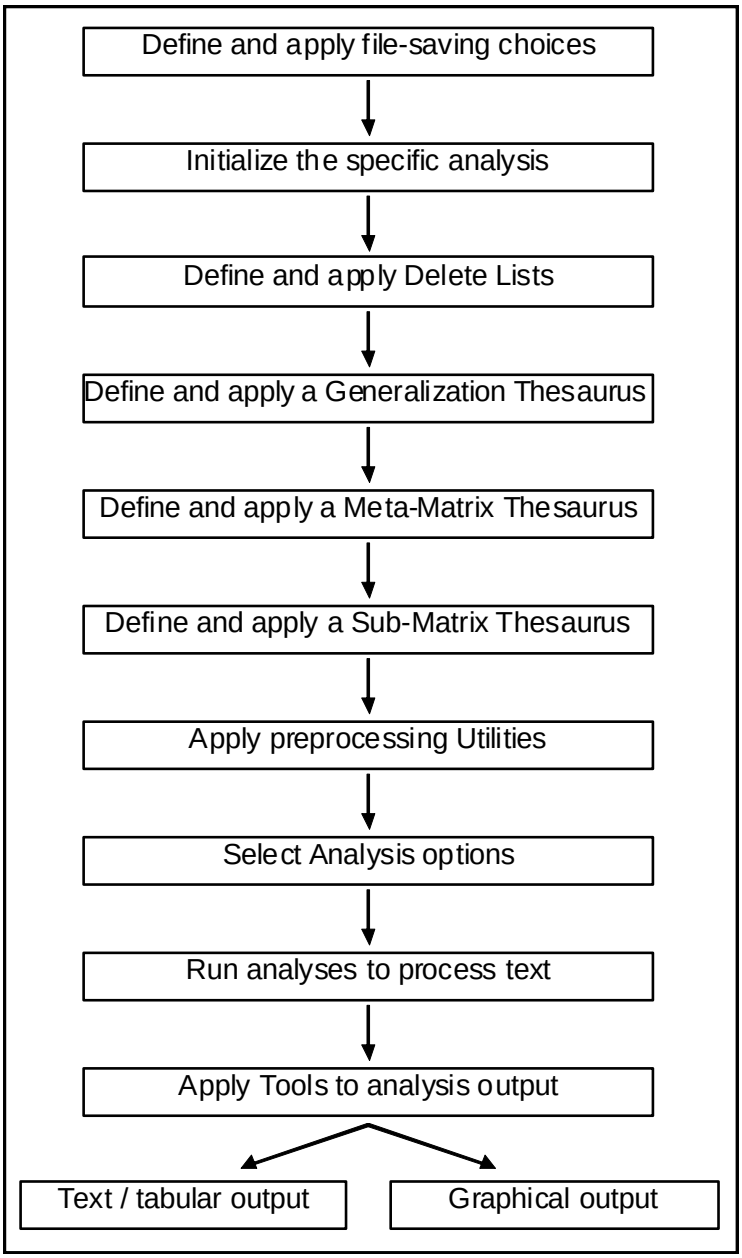


Figure 2. User tasks necessary for use of AutoMap.

5. Define and apply a Meta-Matrix Thesaurus.
6. Define and apply a Sub-Matrix Thesaurus.
7. Apply selected preprocessing Utilities.
8. Select Analysis options.
9. Run analyses to process text.
10. Apply Tools to analysis output.

These are illustrated schematically in figure 2 and discussed in more detail in the next subsections.

The order of application is operator selectable, not determined by the AutoMap program. The implications of several differing application sequences are discussed in more detail where application order is particularly important.

The AutoMap User's Guide states that preprocessing techniques are optional and recommends application, if used, in a specific order. This order differs slightly from that used by the authors. It is recommended in the User's Guide that three utilities be applied first, then the Delete List, etc., as listed previously.* The three utilities—Named Entities Recognition, Collocation/bigram identification, and Stemming—are not generally used by the authors at this time. Named Entities Recognition and Collocation/bigram identification are two of the utilities discussed in section 4.8.†

Stemming is the process of reducing variant forms of certain parts of speech to a common root. For example, "be," "am," "are," "is," and "was" might be reduced to a stem of "is." There is a substantial improvement in brevity and grouping of related events, but the potential loss of meaning is substantial. In particular, temporal and relational or conditional information is lost. These data are crucial for tactical operations. For this reason, stemming is not presently used and is not presented in this report. It may be included in subsequent editions. There is an excellent discussion of stemming and the different techniques for stemming in the AutoMap User's Guide.‡

4.1 Define and Apply File-Saving Choices

The Output Storage Manager

Definition and application of the user's own choices for file names and locations is crucial and should be done immediately. If the user does not do so, the program will store information in the default locations under default file names. Each time the program is opened, the previous file management structure is erased and the default applied. Some files are saved automatically on use, which then erases the prior versions, wasting a great deal of time and possibly leading to gross errors in actual use of the program.

*See section 2.3, "Hierarchy of Pre-Processing Techniques," in the section "Text Pre-Processing" of the Automap User's Guide, accessed from the Help menu of AutoMap, v. 2.6.70.

†A note on terminology: the list of "utilities" is taken from the User's Guide, but the terminology can be confusing. All the items are listed as "utilities," but the first three items are also menu items under a tab specifically named "Utilities" (see [3. Pre-Processing Settings] -> [1. Utilities]). In order to clarify this terminology, the destination menu items from the AutoMap interface are shown with the "utilities" listed.

‡Section 4, "Stemming," of the section "Text Pre-Processing" of the Automap User's Guide, accessed from the Help menu of AutoMap, v. 2.6.70.

An example is the series of “Applied” thesauri shown as choices in the Output Storage Manager screen, shown in figure 3. If the user loads and applies any of the predefined thesauri available through the File Menu, such as the thesaurus for normalizing country abbreviations and names ([File] -> [Open Generalization Thesaurus] -> [Open thesaurus for generalizing countries]), that thesaurus will, on application, overwrite any operator-defined Generalization Thesaurus, losing a great deal of work and possibly leading to the generation of wrong files. As noted in the section on using the Generalization Thesaurus, if the predefined thesauri are applied sequentially, the program undoes each thesaurus when the next one is applied so that an operator may think several modifications to the text have been made, when, in fact, only the latest modification applies. During the process, the operator may have overwritten a thesaurus that took a great deal of time to build.

Further, if the user then analyzes the data without realizing that the output actually applies to another input file or set of files, the result is disaster.

To save data, select [File] -> [Output Storage Manager]. This will produce a screen where one can change the directories and filenames in which the data will be stored.

Notes: • You must change the directories and/or filenames or previous data will be overwritten and not saved. It is extremely important to do this for each analysis. Unless the destinations for each output are selected by the operator, the software will use the default settings and overwrite all previous results.

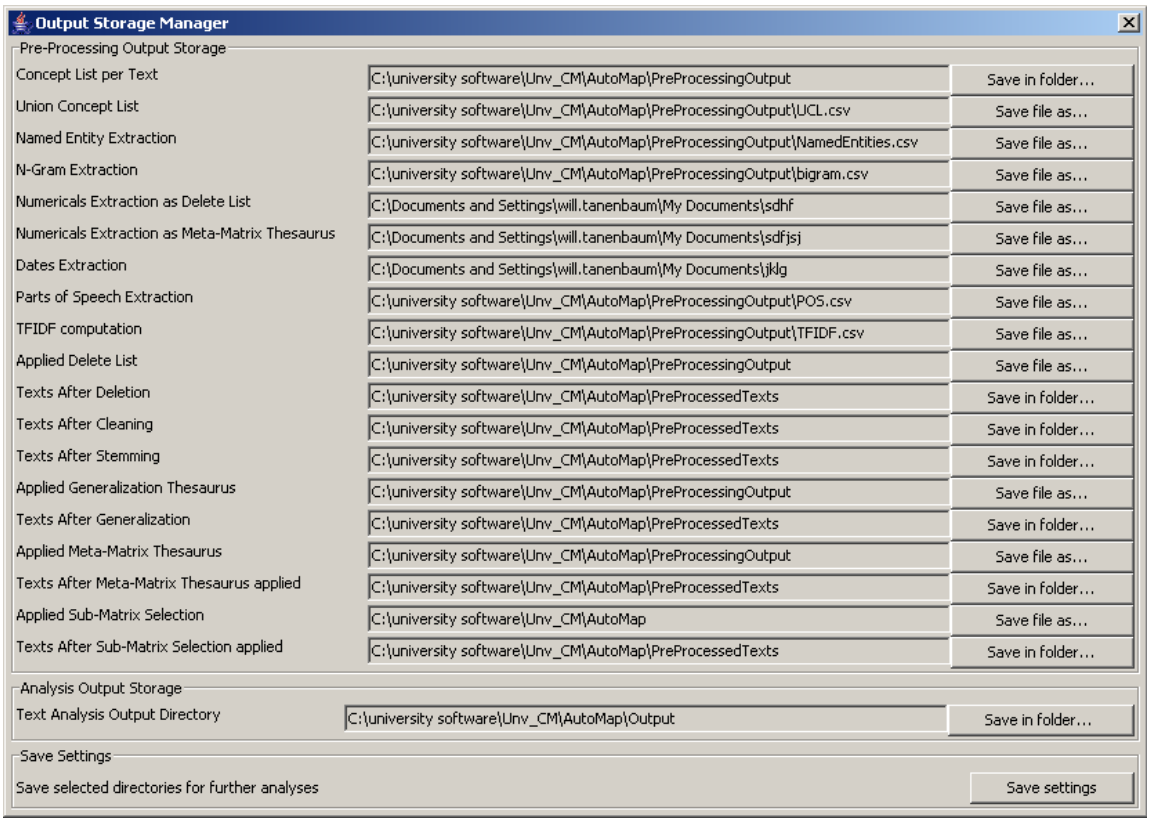


Figure 3. The Output Storage Manager screen.

- These changes will not be stored and must be remade upon reentry into AutoMap.
- It is advisable to keep a saved list of directory paths for the Output Storage Manager for each analysis.

4.2 Initialize the Specific Analysis

Each analysis my have a different target set of text file inputs and require a different set of preprocessing choices. In some cases, a single target set may be analyzed several different ways for different information, requiring different preprocessing choices.

4.2.1 Load Input File(s)

The first task that must be accomplished after selecting file storage names and locations is to load the input file or files. AutoMap can analyze single texts or a family of texts located in a common directory. Begin by opening the document or documents to be analyzed. Select [File] -> [Open Single File], select path to folder or [File] -> [Open Multiple Files {select folder}], select path to folder.

4.2.2 Concept Lists

A concept is a word, words, number, or symbol from the document one is examining; thus the Concept List is the list of all concepts from the document under analysis. If the User has loaded multiple texts, the document being preprocessed will be shown in the upper-left window of the screen shot in figure 3.

A key basis for analysis is the Concept List. It is generated upon loading the input file(s). The Concept List shown in the Concept List Window, activated by the Concept List tab, displays the Concept List for the document shown in the Original Texts window. That is, the Concept List applies to whatever document is active.

The Union Concept List, shown by selecting the Union Concept List tab, is the list of all the concepts in the complete file of documents that has been loaded. If only one document has been loaded, the Concept List and Union Concept List are the same.

The Union Concept List must be generated by using the file menu. It may be created or refreshed at any time but will show the Delete List only if created after the Delete List is created, applied, and unapplied. Once the Delete List has been created, select [File]-> [Create and refresh Union Concept List]. The Concept List or Union Concept List can be viewed by selecting either [1. Concept List] or [2. Union Concept list]. These tabs and the Concept List are shown in figure 4.*

*The document shown in the figure 4 screen shot was supplied by Dr. Michelle Vanni, U.S. Army Research Laboratory. Dr. Vanni provided several sets of original (source) documents and seven translations of the source. Each set included a short passage in Spanish and seven translations—two human and five machine translations. Material provided by Dr. Vanni is found in figures 6, 10, and 13. These are referred to as various translations of documents numbered 6, 7, or 8. Other document or text excerpts are invented by the authors.

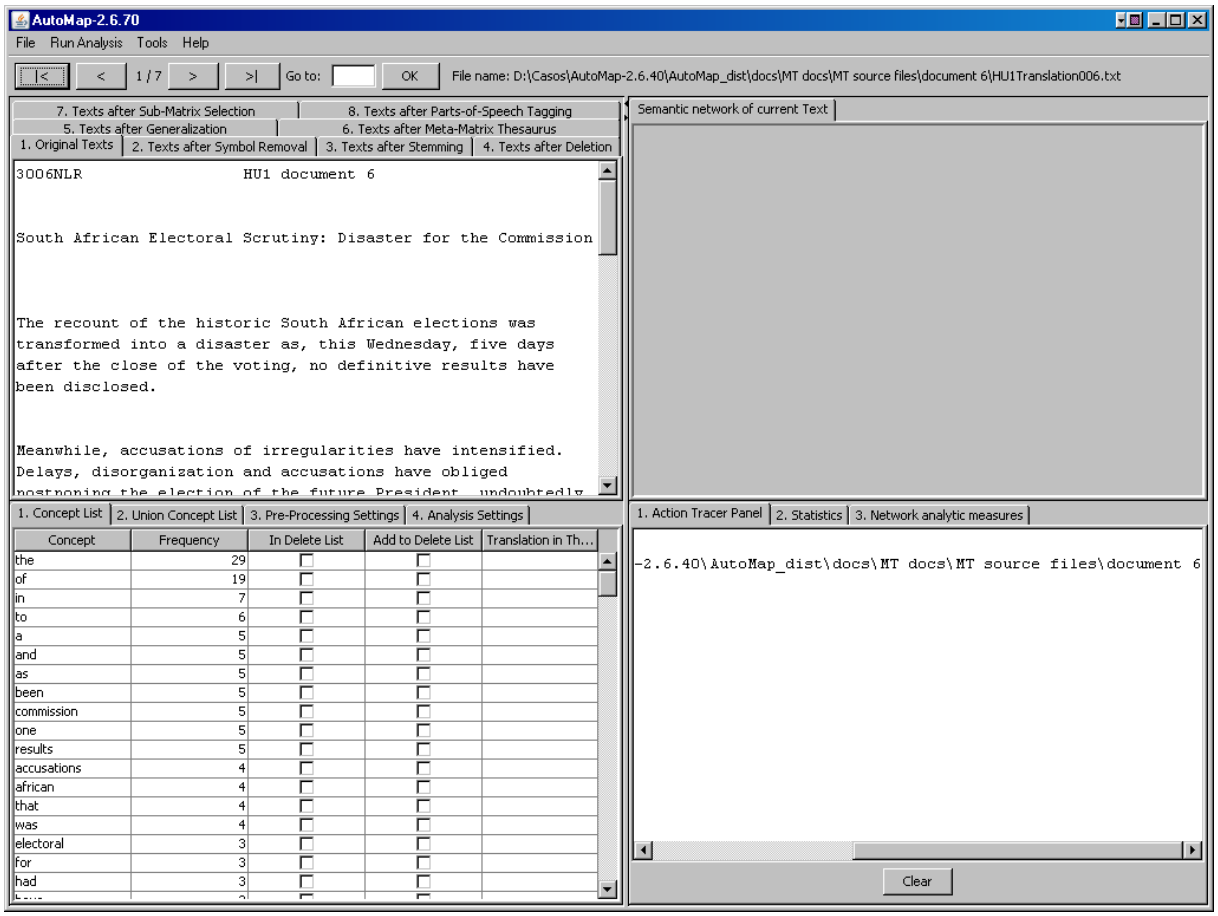


Figure 4. Text files loaded for analysis. The specific text is shown in the upper-left window, with the numerical sequence number shown in the array of buttons below the menu bar.

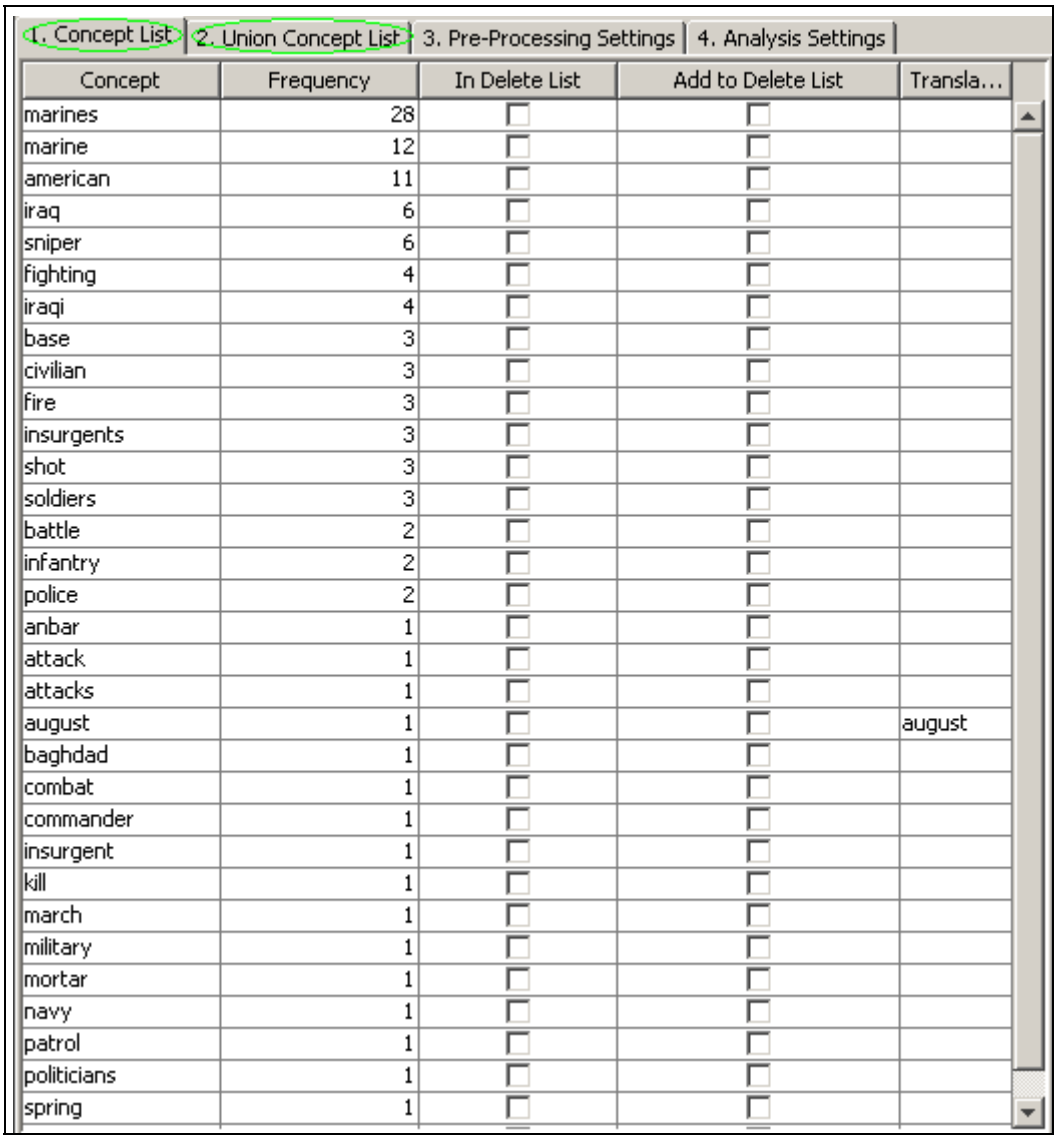
Note: The Concept List is altered by the Delete List, Generalization Thesaurus, and Meta-Matrix Thesaurus.

A Union Concept List is shown in figure 5. This Union Concept List was created after the Delete List was created, applied, and unapplied (see next section). The concepts in the Delete List are shown as checked, the result of the application-unapplication sequence.

4.3 Define and Apply Delete Lists

Two ways to obtain a Delete List are as follows:

1. Loading a premade Delete List. Select [File] -> [Open Delete List] -> [Open from file].
2. Creating a Delete List. A Delete List is a list of words the user defines to be removed from the text. The Delete List can be viewed by selecting [3. Pre-Processing Settings] -> [3. Delete List].



Concept	Frequency	In Delete List	Add to Delete List	Transla...
marines	28	☐	☐	
marine	12	☐	☐	
american	11	☐	☐	
iraq	6	☐	☐	
sniper	6	☐	☐	
fighting	4	☐	☐	
iraqi	4	☐	☐	
base	3	☐	☐	
civilian	3	☐	☐	
fire	3	☐	☐	
insurgents	3	☐	☐	
shot	3	☐	☐	
soldiers	3	☐	☐	
battle	2	☐	☐	
infantry	2	☐	☐	
police	2	☐	☐	
anbar	1	☐	☐	
attack	1	☐	☐	
attacks	1	☐	☐	
august	1	☐	☐	august
baghdad	1	☐	☐	
combat	1	☐	☐	
commander	1	☐	☐	
insurgent	1	☐	☐	
kill	1	☐	☐	
march	1	☐	☐	
military	1	☐	☐	
mortar	1	☐	☐	
navy	1	☐	☐	
patrol	1	☐	☐	
politicians	1	☐	☐	
spring	1	☐	☐	

Figure 5. The Concept List screen. A Concept List applies to a single document. The Union Concept List tab will display all the concepts from a series of documents considered and analyzed together.

By clicking inside the window, one is able to position a cursor to type words into the list (one word per line). Selecting [Apply Delete List] will remove those words from the text. Certain punctuation marks are registered as parts of words. For instance, the term &Fred must be entered into the Delete List as &Fred instead of simply Fred, and vice versa. A word's entry into the Delete List is entirely up to the user. The fewer words in a text, the less cluttered a document becomes, but over-deletion can cause a critical loss of key information.

A Delete List may also be constructed as a .txt file by use of a word processor. The desired list is then loaded just like any predefined list or sections inserted into the Delete List panel by the Paste function of Windows.

One advantage of using the Windows processing tools is alphabetization. A circuitous application of Excel and an application such as Notepad allows alphabetization in Excel and detection of duplicate entries. Duplicate entries do no harm, but it is sometimes useful to list all similar forms and examine them for completeness. An example is the set of variant transliterations of the Arabic word "al." A name such as "al Masri" may be translated as "al-Masri," "al Masri," or "alMasri." It is necessary to insert the undesirable variant forms into the Delete List. The alphabetization function has not, as of this writing, been incorporated into the Delete List panel or, for that matter, the Generalization Thesaurus panel. A similar set of operations is useful in building a Generalization Thesaurus. In that case, listing all variant spellings of a word allows standardization.

Figure 6 illustrates how terms in the Union Concept List are indicated to be in the Delete List. This screen is useful for determining whether some words that could be deleted have not yet been chosen for deletion. In this case the term "results" may be considered for deletion, depending on the context and intent of the analysis.

Figure 7 shows the window used for management of the Delete List; no concepts have yet been entered. As can be seen, a delete list may be applied and "un-applied." This is useful if it is decided to restore a word previously deleted. That is, a Delete List may be "un-applied," restoring those terms to the text. A concept may then be removed from the Delete List and the list applied. This restores the desired concept to all locations in the text where it belongs.

Notes:
- Words must be typed in all lowercase. Even if they are capitalized in the article, they must be in lowercase or they will not be removed.
- Beware of special characters; they turn up frequently. Copy and paste from the article into the Delete List if found. The only way to determine if a special character is present is when a word is not deleted even though it is in the Delete List.

One can also create a Delete List by selecting [1. Concept List] and selecting the concept in the [Add to Delete List] column. (This is unadvisable because the interface often requires clicking upwards of six times before it registers.) The Delete List indicator is shown in the Concept List panel illustrated in figure 8.

One can also create a Delete List by clicking on [2. Union Concept List] and clicking on [Concepts] (words) in the Add to Delete List column. (This is also unadvisable because adding further concepts to a prewritten Delete List will erase the prewritten list.)

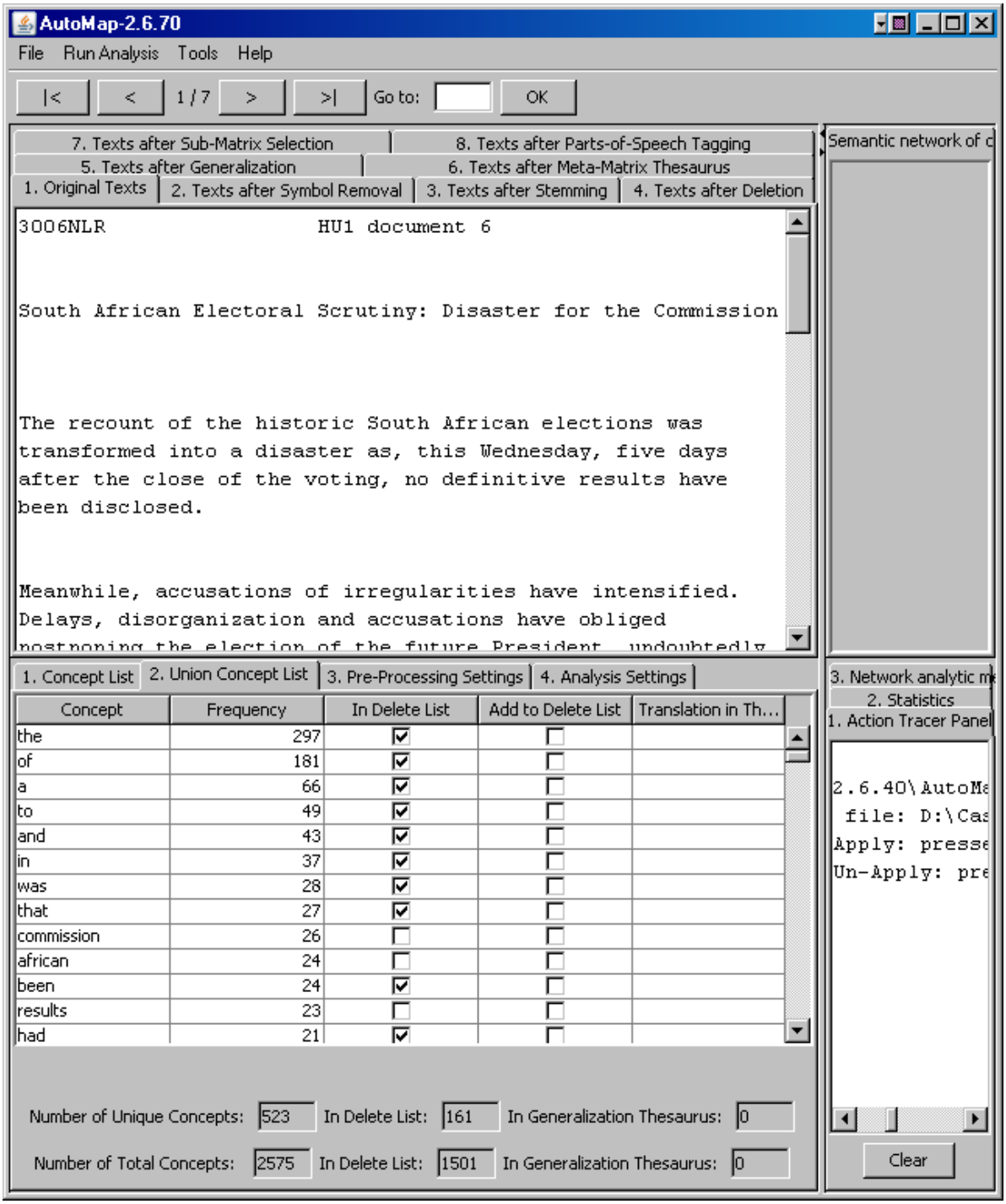


Figure 6. The Union Concept List. Note the concepts in the Delete List are shown.

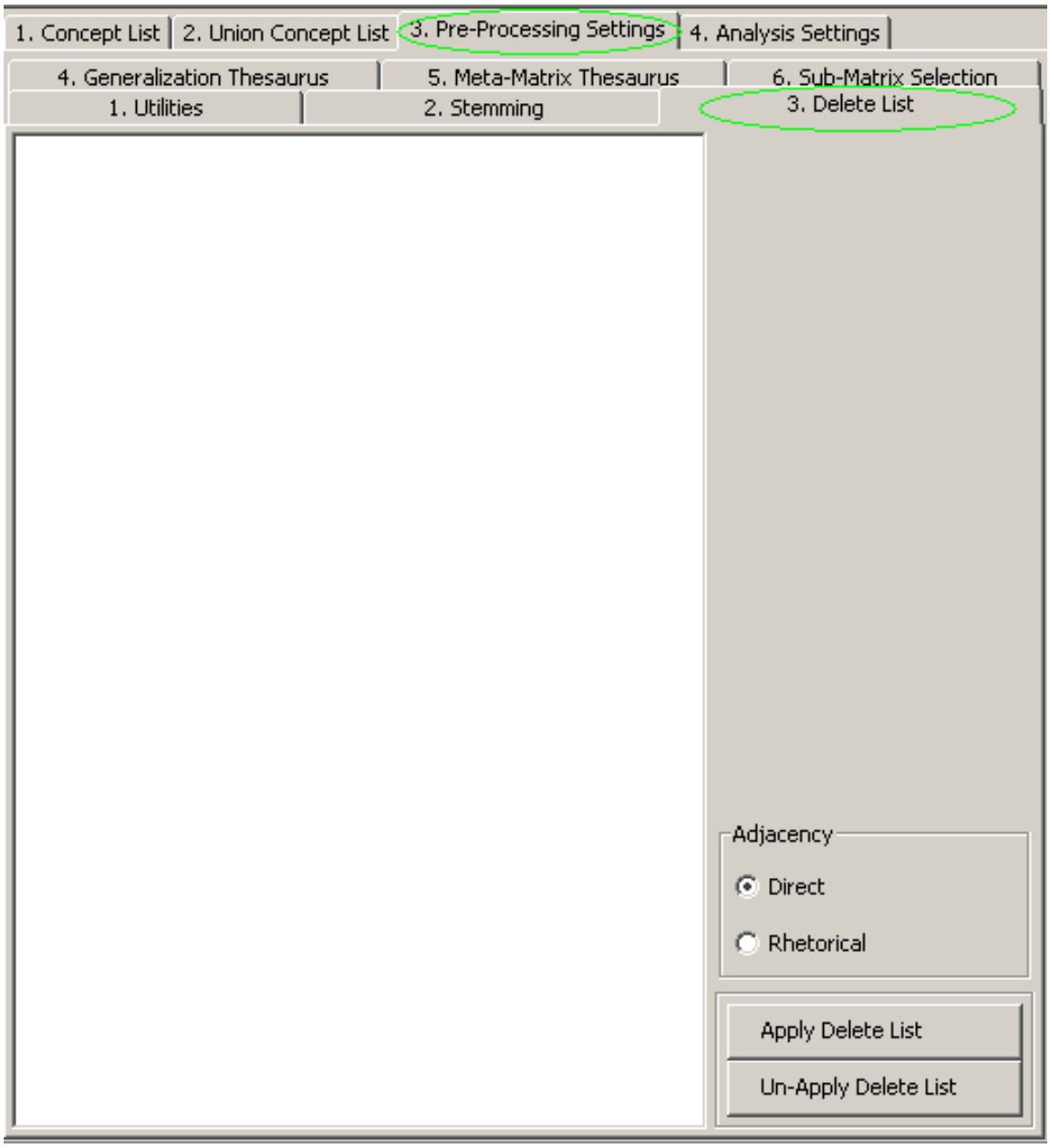


Figure 7. The Delete List panel. No Delete List has been loaded or constructed yet.

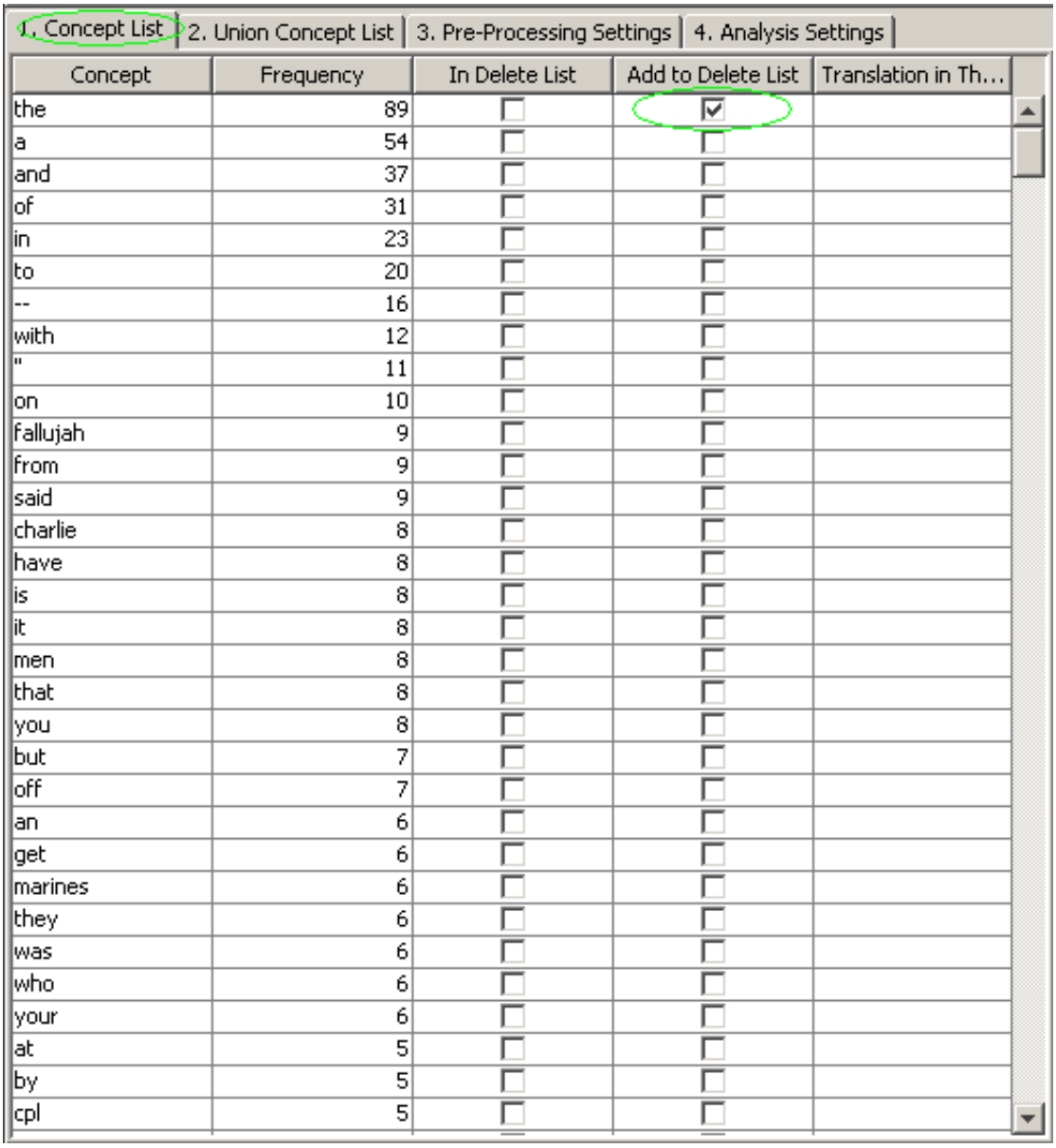


Figure 8. Delete List entries in the Concept List panel.

4.4 Define and Apply a Generalization Thesaurus

As described earlier, a Generalization Thesaurus is a means of linking multiple words or symbols into a text string that can be analyzed as a single concept. As with the Delete List, one may use a predefined list as a Generalization Thesaurus or create one.

4.4.1 Loading a Premade Generalization Thesaurus

Select [File] -> [Open Generalization Thesaurus] -> [Open from file thes].*

* "Thes" is the actual caption in the user interface, not a misprint in this report.

4.4.2 Creating a Generalization Thesaurus

A Generalization Thesaurus is words in a list that are identified as the same thing. An entry can be a name, an infinitive, a single word, or any other two words labeled as the same concept. The Generalization Thesaurus can be viewed by selecting [3. Pre-Processing Settings] -> [4. Generalization Thesaurus]. Figure 9 shows the panel that controls the Generalization Thesaurus.

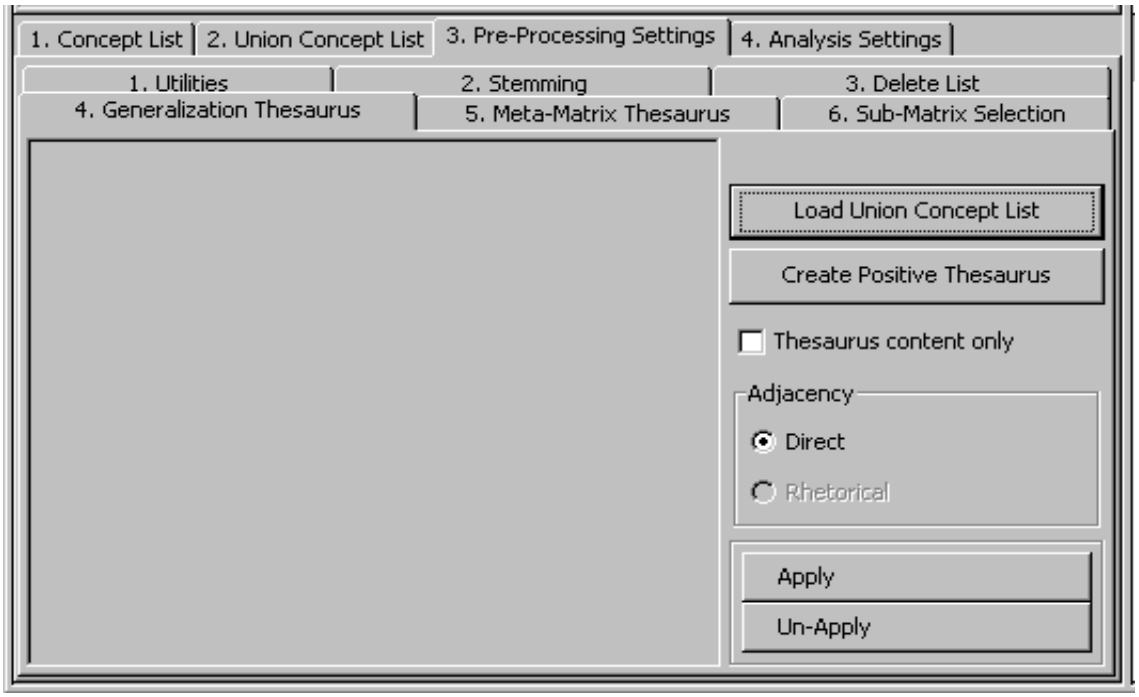


Figure 9. The Generalization Thesaurus panel. No Generalization Thesaurus has been loaded or entered.

To create a new thesaurus in the AutoMap window, load a blank text file as a Generalization Thesaurus through the [File] menu. This will open the panel for use. The modified thesaurus must be saved as such (see saving files). By clicking in the window, the cursor is positioned so that one may type entries into the thesaurus (one entry per line). An entry is a word or words separated by a comma such that the first term is changed into the second term. For example, the term KwaZulu-Natal is converted to KwaZulu_Natal.

Selecting [Apply] will activate the Generalization Thesaurus generation tool.

The use of the Generalization Thesaurus is like that of the Delete List; it is not required, but its use removes clutter and increases document clarity. Figure 10 shows the form of a Generalization Thesaurus in the Generalization Thesaurus control panel. Note each entry is a comma-delimited pair, with the term to be modified on the left and the modified term on the right. This ability is especially important when several variants of the same idea or concept may be in simultaneous use.

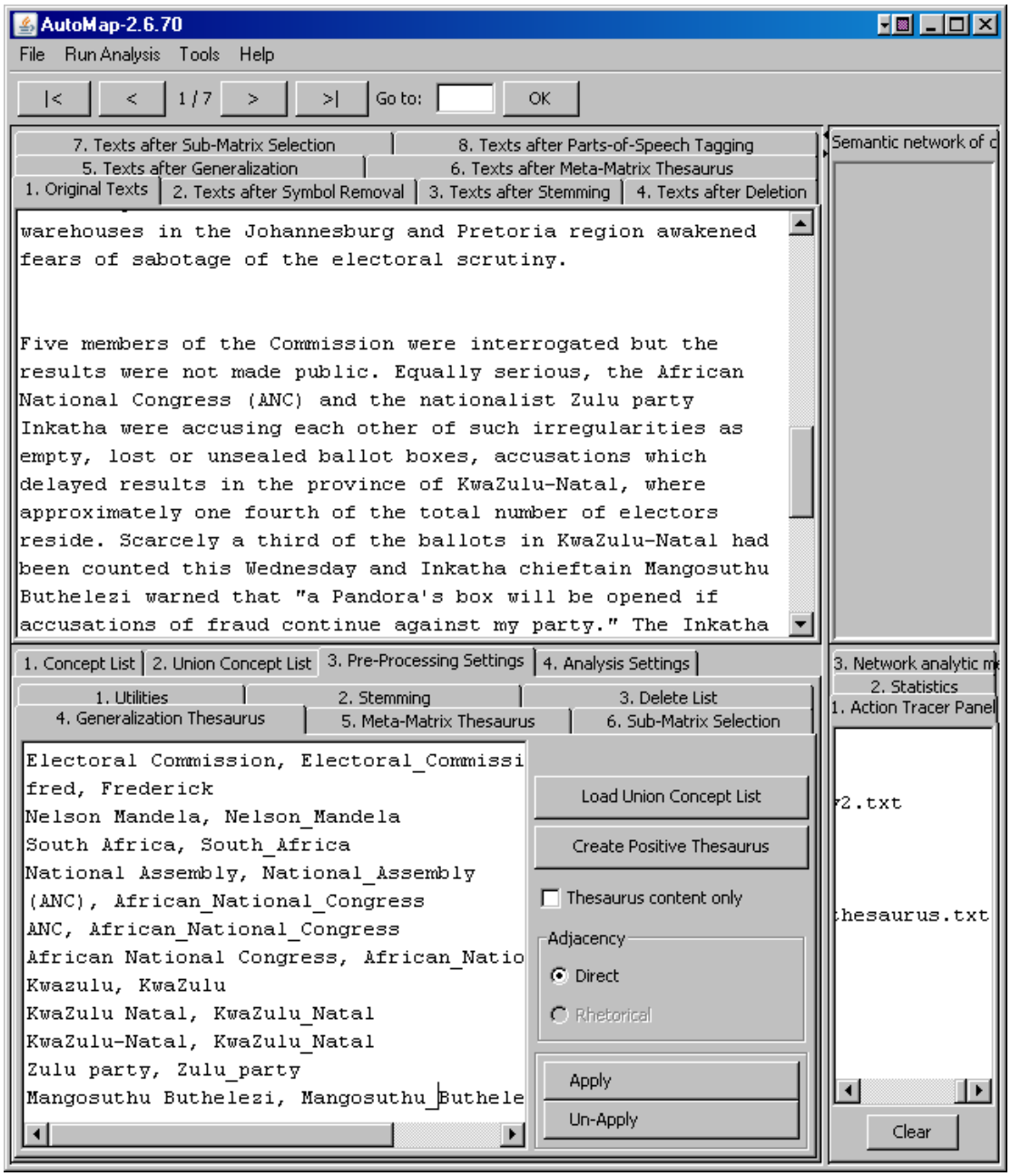


Figure 10. A Generalization Thesaurus has been created. Note the original text in the upper-left window is still unchanged, as it should be.

Examples might be the use of an acronym to indicate a thing (“ARL”) and more formal terms (“Army Research Lab” and “Army Research Laboratory”), along with incomplete forms (“the lab”) to indicate the same basic idea. This is especially important when several transliterations or naming conventions may be in simultaneous use (“al-Zawahiri,” “al Zawahiri,” and “alZawahiri,” etc.). An analyst must use care as it would be easy to insert meaning where none or different meaning was intended.

A Generalization Thesaurus may also be created off-line in a text processor such as Notepad. Additionally, material from Notepad can be pasted into the AutoMap Generalization Thesaurus panel. Alphabetization and checking for duplicates and completeness can be accomplished by use of Excel and Notepad, as described in the Delete List section.

Note: Entries need not be case sensitive; however, use a comma followed by a single space to separate one concept from another.

4.5 Define and Apply a Meta-Matrix Thesaurus

A Meta-Matrix Thesaurus is a classifications list of all the concepts left in one's document after its preprocessing. A concept can be classified as a knowledge, agent, resource, task, event, organization, location, role, time, attribute, or user-defined. A Meta-Matrix Thesaurus may be created or a predefined one loaded.

4.5.1 Loading a Predefined Meta-Matrix Thesaurus

Select [File] -> [Open Meta-Matrix Thesaurus] -> [Open from file], then navigate to and select the file.

4.5.2 Creating a Meta-Matrix Thesaurus

To create a Meta-Matrix Thesaurus select [File] -> [Meta-Matrix Thesaurus] -> [Open from highest level of pre-processing]. The Meta-Matrix Thesaurus screen can be viewed by selecting [3. Pre-Processing Settings] -> [5. Meta-Matrix Thesaurus]. The concepts that remain in the document after that level of preprocessing will be displayed. By selecting the desired column or columns, a concept may be labeled as any of the preloaded classifications (agent, event, etc.) or user-defined labels may be created.

Alternatively, one can create a Meta-Matrix Thesaurus in a word processor and manually type the thesaurus. An example of two entries in such a thesaurus is

1. cover, task
2. prime_minister_nouri_al-maliki , agent , role.

In the first entry, the term "cover" is classified as a "task." In the second, the concept "prime_minister_nouri_al-maliki" is assigned the classifications "agent" and "role," Note this is a comma-delimited list format.

Selecting [Apply] will activate the Meta-Matrix Thesaurus and apply it to the document.

Notes:
- Opening a Meta-Matrix Thesaurus from the highest level of preprocessing erases any thesaurus you currently have loaded and applied.
- The only way to augment a Meta-Matrix Thesaurus with new concepts is to modify the file directly by using a word processor.

- The Meta-Matrix Thesaurus is not optional. Without classification, your concepts are meaningless to outside programs such as Organization Risk Analyzer (ORA) or Starlight.
- The more classifications you assign to a concept, the more cluttered its depiction will be.
- The fewer classifications you give a concept, the greater the risk of missing a key connection between concepts.

4.6 Define and Apply a Sub-Matrix Selection

4.6.1 The Sub-Matrix Selection

The Sub-Matrix Selection defines the manner in which concepts, with roles or functions identified through the Meta-Matrix Thesaurus, are linked. The Sub-Matrix Selection can be accessed by going to [3. Pre-Processing Settings]-> [6. Sub-Matrix Selection]. Select [Add new line] to begin, then select [Add in same line] to connect categories to the original. An example may be found in figure 11. Note that this option cannot function unless the Meta-Matrix has been defined and applied.

4.6.2 Loading a Premade Sub-Matrix Selection

To load a predefined Sub-Matrix Selection, select [File] -> [Open Sub-Matrix Selection] -> [Open from file].

4.6.3 Loading the Full Meta-Matrix Thesaurus

A Meta-Matrix Thesaurus is shown in figure 12. In this screen, the roles of the terms shown are defined. For instance, the term crimi[nal] may have the functions of an “agent,” an “organization,” or a “role.” Examples of those functions might correspond to usages such as “a criminal escaped,” “criminal(s),” and “criminal gang.”

The tool for managing a Sub-Matrix is shown in figure 13.

To load the Full Meta-Matrix Thesaurus, select [File] -> [Open Sub-Matrix Selection] -> [Load Full Meta-Matrix]. The Full Sub-Matrix shows the connectivity between every classification category.

Selecting [Apply] will activate the Sub-Matrix Selection.

Note: Without a Sub-Matrix Selection, the visualization of the Meta-Matrix Thesaurus will appear as singular-floating concepts with no statement. This is described in the section on Analysis Tools.

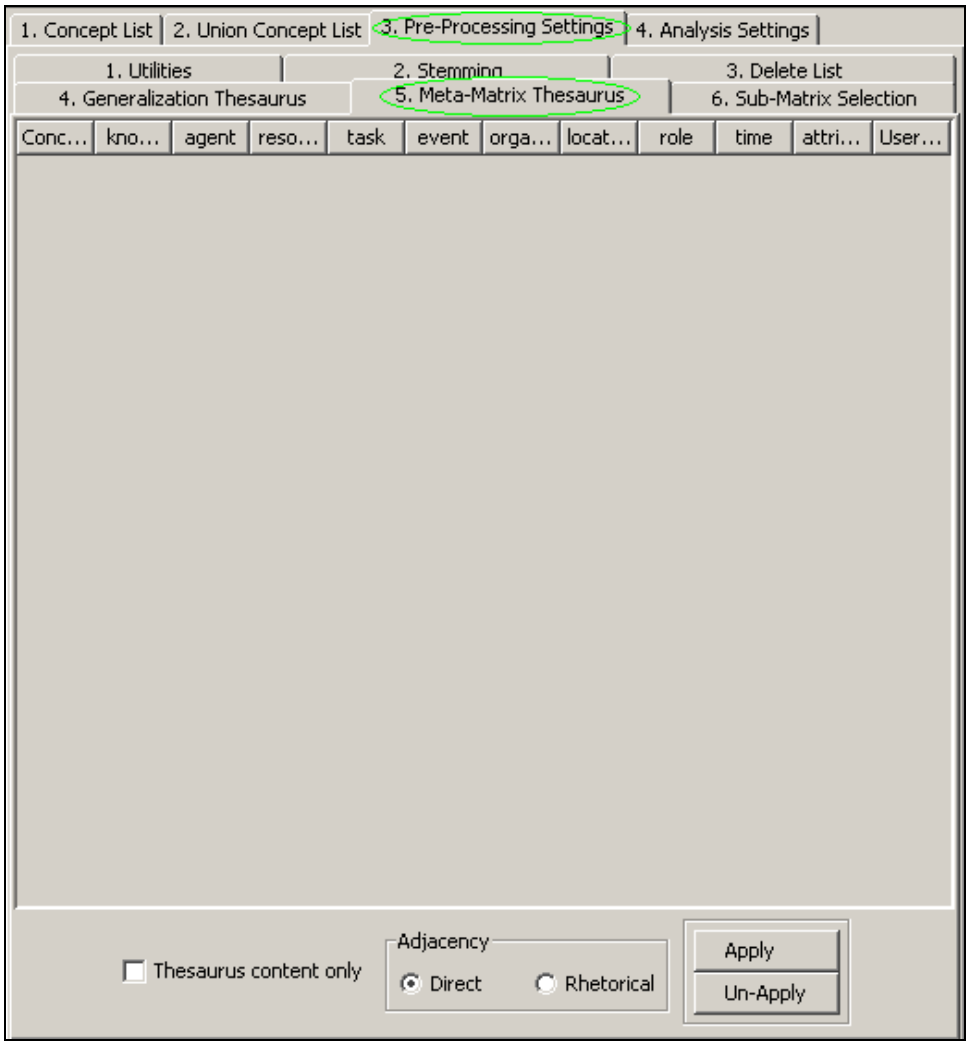


Figure 11. The Meta-Matrix panel of the AutoMap interface.

4.7 Apply Preprocessing Utilities

4.7.1 Cleaning up Texts

Utilities that "clean up" texts can significantly reduce distractions to seeing and understanding key patterns or relationships. This may be extremely useful if the analysis does not require use of the characters removed by the cleanup utilities. The different utilities are shown in figure 14. The utilities must be used carefully as it is easy to remove key data. For example, dates may be extremely important; the association of "Normandy" and "6 June 1944" are cases of the past importance of such relationships. Another might be "9/11/01" and "New York." Special characters such as & or @ may be very important as well (e.g., "osama.bin.laden@some .isp.com" or "B&B"). The example "27 Dhubat Street" illustrates the potential impact at the tactical level. If one is to search for a bomb hidden on Dhubat Street, it helps to know which door to knock on.

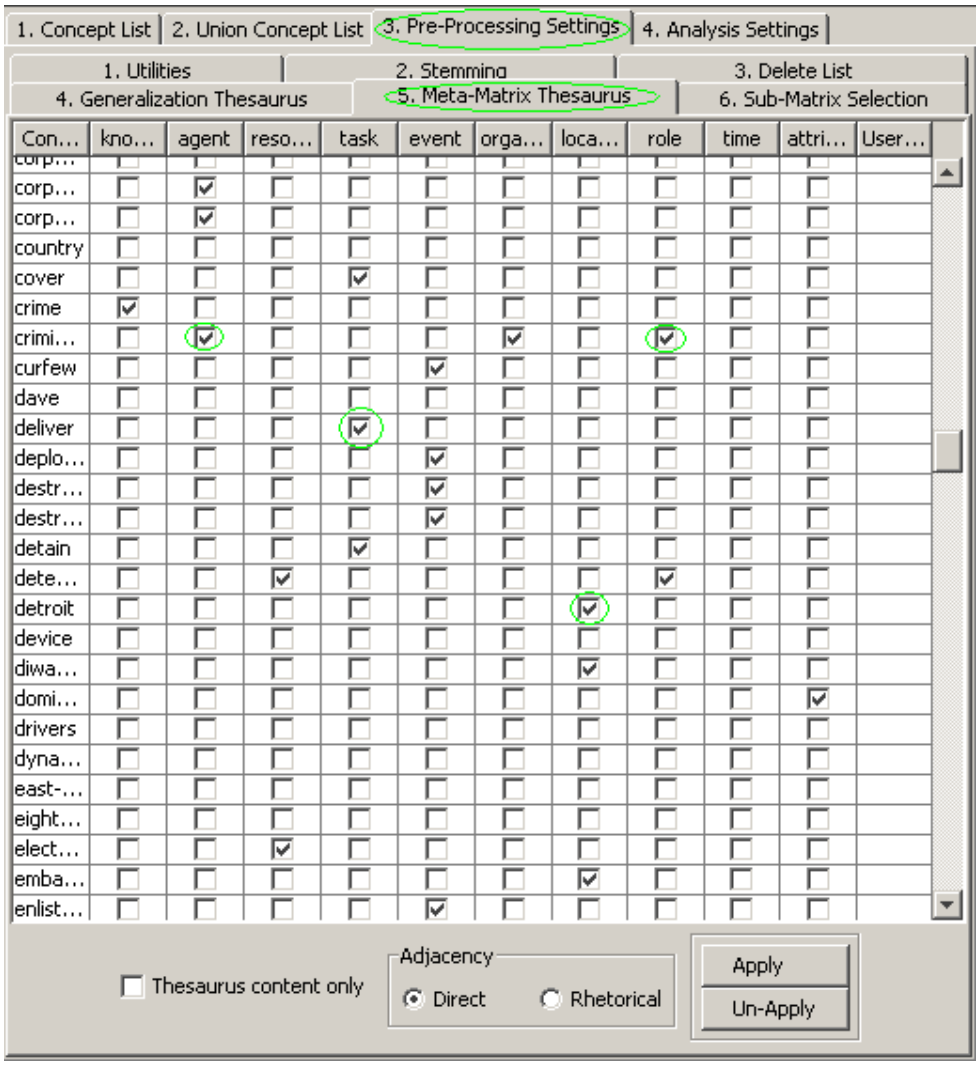


Figure 12. An example of a Meta-Matrix Thesaurus.

It has been found useful to delay removing symbols and numbers until the analysis has taken shape. At that time, the option may or may not be exercised. This action is performed by going to [3. Pre-Processing Settings]-> [1. Utilities] -> [Remove Symbols] or [Remove Symbols and Numbers].

Note: Alternatively, one can manually add the unneeded and undesirable symbols and numbers to the delete list and avoid losing any desired data.

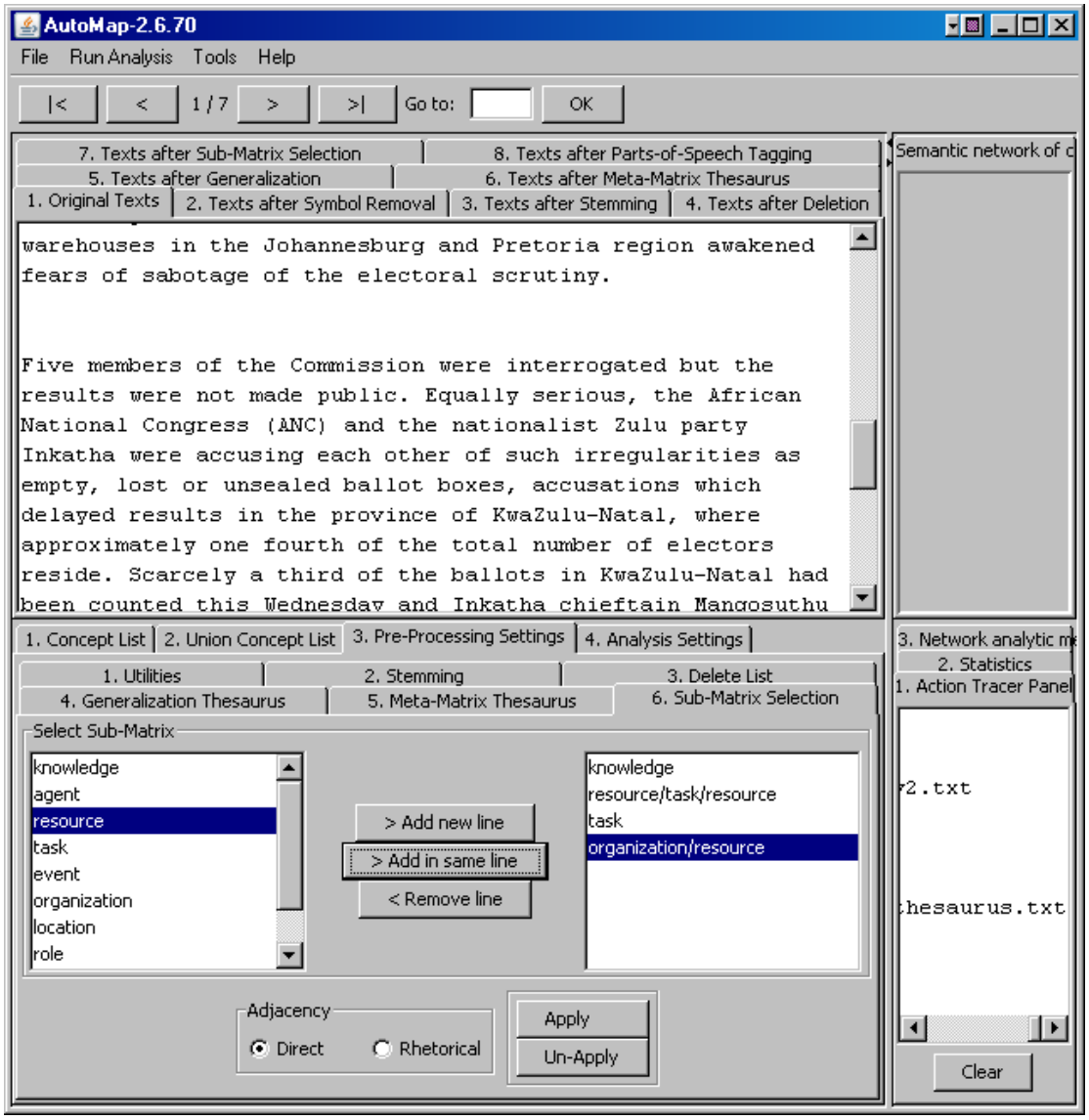


Figure 13. The Sub-Matrix control panel.

4.7.2 Parts of Speech Tagging

Parts of Speech Tagging produces a file with the part of speech estimated by the software for each concept. This option has not been useful in the analyses conducted so far.

4.7.3 Named Entity Recognition

Named Entity Recognition produces a list of concepts believed by the software to be names. This list has so far been full of extraneous concepts but may be useful.

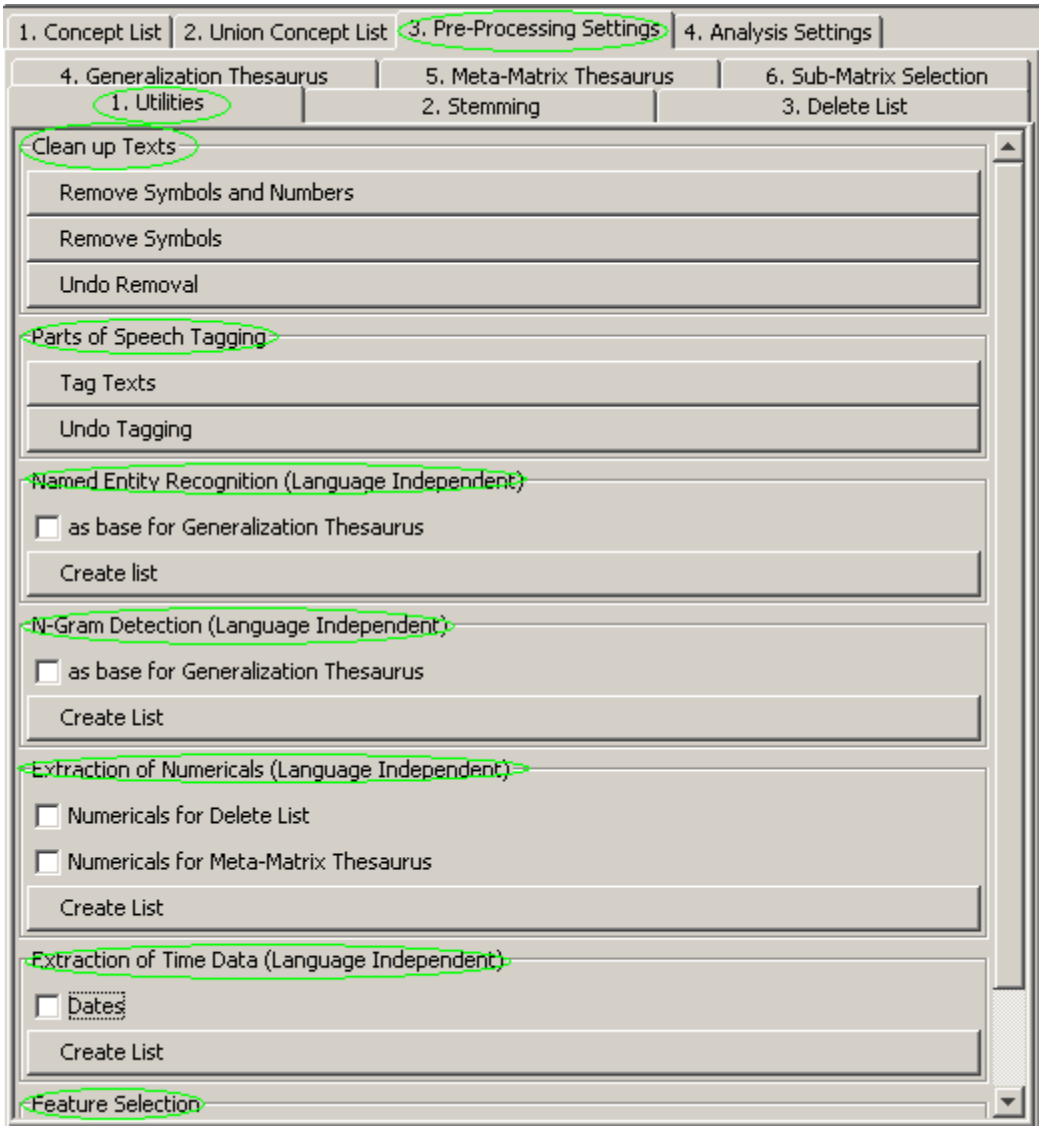


Figure 14. The Utilities panel.

4.7.4 N-Gram Detection

N-Gram Detection creates a list of concept pairs, or bigrams. This set of pairs may aid in constructing the Generalization Thesaurus. For example, the terms “abu” and “Khalid” may appear as juxtaposed pairs quite often; that may indicate that the terms are actually linked to the same concept, represented in the source text as a multiple word concept. The translation of “abu Khalid” to “abu_khalid” may be an appropriate entry for the Generalization Thesaurus. On the other hand, the juxtaposition of “Achmed” and “Dhubat” probably refers to a link between the individual “Achmed” and the location “Dhubat.” The N-Gram list must therefore be considered both a powerful shortcut to meaning and a great way to err badly and quickly.

4.7.5 Extraction of Numericals

Extraction of Numericals creates a list of all numerals to add to the Delete List or possible concepts with numerals to add to the Meta-Matrix Thesaurus. Note that actually adding the numerals to the Delete List must be done manually. It must be emphasized that this action may reduce clutter but may also destroy vital meaning. It is potentially dangerous.

Note: This function cannot be performed after any other preprocessing step.

4.7.6 Extraction of Time Data

Extraction of Time Data creates a list of some of the dates in the document. This may be convenient for analyses focusing on associations with a key date.

4.7.7 Feature Selection

Feature Selection creates a TF-IDF (term frequency–inverse document frequency) weight. This is an aid to filtering out noise from concepts that are frequent but not substantive by highlighting and differentiating concepts used relatively sparingly. This is not always a key to finding important relationships but may under some circumstances be a useful aid in first-time analyses of a given situation. For instance, if a team has entered an entirely new area with entirely new tribal affiliations, a different set of concepts or terms will be used frequently compared to the previous situations in previous areas of operations. As a hypothetical example, in a new area the tribal affiliation "osmanli" might be dominant; in the previous area of operations the dominant tribal affiliation term might have been "fatimid." A report of a first contact between the "osmanlis" and the "fatimids" would highlight the term "osmanli" with a higher TF-IDF, flagging it for examination and so discovering a very important phenomenon.

The term frequency, *tf*, is a weighted measure of the number of times a given term occurs in a document. A relatively unimportant term such as "the" may occur many times in a document, and another term which is important may occur only a few times, yet be a critical part of the meaning of the whole text file. Likewise, a term may occur only once and yet be completely unimportant. A balance between frequency and importance is needed.

A measure that is used to balance frequency and importance is the term frequency-inverse document frequency, or tf-idf.[*] Thus

[*] The measure term frequency, or *tf*, is the frequency of a word used in a document. An interesting note is that different authors use different definitions. For example, some references found in a cursory search on the search term *tf-idf* use for a definition of term frequency a *term count*, or number of occurrences of a word in a document, and others an actual *term frequency*, or number of occurrences divided by the total number of terms in the document. Introducing *term weight* also gives an interesting variety of definitions, but term weight is not used here. This report uses the *frequency* version of the term. An example of frequency can be found at http://www.stanford.edu/class/cs276/handouts/lecture6-tfidf.pdf, accessed 20 March 2008. An example of term count used as term frequency may be found at http://archimedes.fas.harvard.edu/presentations/2002-03-09/img13.html, accessed 20 March 2008, or Gyongyi and Garcia-Molina's "Web Spam Taxonomy" at http://airweb.cse.lehigh.edu/2005/gyongyi.pdf, accessed 20 March 2008.

$$tf_{i,j} = \frac{n_{i,j}}{m_j}, \quad (1)$$

where $Ni_{,j}$ is the number of occurrences of the *ith* term in the document d_j, and m_j is the total number of all terms in the document d_j.

The inverse document frequency is a way of assessing the importance of a term in the whole body of *m* documents under analysis, or corpus.

$$idf_i = \log \frac{m}{df_i}, \quad (2)$$

where *m* is the total number of documents in the document set or corpus, and df_i is the number of documents containing the *ith* term. The tf-idf is thus

$$tf - idf = tf_{i,j} * id_f, \quad (3)$$

or

$$tf - idf = tf_{i,j} * \log \frac{m}{df_i}. \quad (4)$$

4.8 Select Analysis Options

4.8.1 Analysis Settings

The Analysis Settings determine the nature of the output information AutoMap produces, as do the Output Options. To access these options, go to [4. Analysis Settings]-> [1. Analysis Settings] or [2. Output Options]. In [Analysis Settings] one may change window size to [4], and in [Additional Output Options], under [UciNet] and the two [DynetML] options, select [Maps], [Term Distribution Matrices], [per Maps], [per TextSet], and [per Text].

4.8.2 Analyses Levels

There are three levels of analysis available to the user—Map Analysis, Meta-Matrix Analysis, and Sub-Matrix Analysis. These are further subdivided by whether single texts or groups of texts are to be processed. These options are straightforward and are well-discussed in the User's Guide, so this subject will not be elaborated in this report.*

*See the section "Analyses" in the AutoMap User's Guide, accessed through the help menu.

4.9 Run Analyses to Process Text

Run the analyses of the documents under investigation after the preprocessing is finished by selecting [Run Analysis] -> [Multiple Map Analysis], then [Run Analysis] -> [Multiple Meta Matrix Text Analysis], then [Run Analysis] -> [Multiple Sub Matrix Text Analysis]. This will create input files that can be used in the associated data visualizers.

4.10 Apply Tools to Analysis Output

4.10.1 Associated Tools

There are tools included in the CASOS ensemble that are associated with AutoMap. One important tool is ORA. ORA is a powerful asset, and its extensive user's guide is full of practical examples. In general, the ORA displays and generates statistical measures describing arrays of entities that can be represented as networks. The relationships between entities can be analyzed by selecting perceptual differences between the glyphs representing entities and groups of related entities in terms of position or graphical means, such as color or shape. The program allows the user to select predefined relationships, such as input Meta-Matrix labels ("organization," "knowledge," "action," "agent," etc.), to define relationships manually by instrument panel entries or by statistical relationships discovered by analysis within the ORA program.

Output can be either textual, graphical network displays, or numerical graphs. The input panel for ORA is shown in figure 15. An example of the textual report is shown in figure 16, the graphical network display in figure 17, and the graphs in figure 18.

ORA is an extremely versatile tool and is updated constantly. The examples shown are from ORA v. 1.8.5. The initial default display shows the nodes "crawling around" the display. The nodes can be selected and dragged to any part of the screen to allow the analyst to group relationships. The directionality of the relationships, if preserved during the analysis in AutoMap, can also be shown. In this way, the difference between "man bites dog" and "dog bites man" is preserved and enhanced. There is a host of options available for font size, map size, and the like, but these will not be enumerated. There are several display options that will analyze and display relationships.

There is a substantial tool set for analyzing the graphs. These will be of considerable interest as operational scenarios and techniques are generated. Clear examples are the relationship between entities and the area of influence about a given entity. These are among the information generated and displayed using the Analysis menu in ORA. Select [Analysis] -> [Generate Reports]; this generates a Selection Wizard. In this example, the option [Who] was selected in the Selection Wizard, which generated a series of selection screens. The report option selected

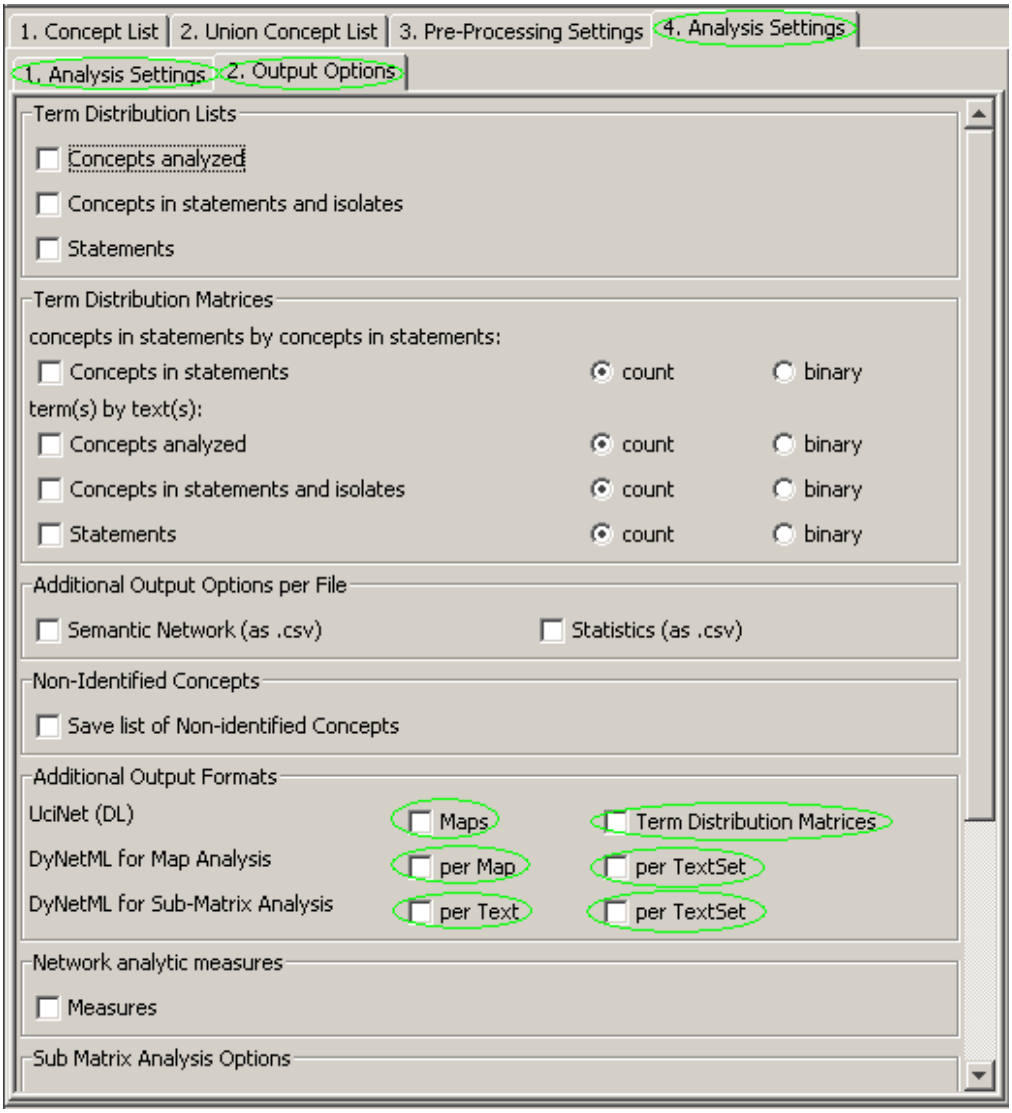


Figure 15. The Output Options screen.

was […path between actors…] then [Next] with "council_of_state" chosen as Entity 1 and "ruud_lubbers" chosen as Entity 2. (The common and inappropriate entity class "knowledge" is an artifact of the files chosen from the original AutoMap output.) The reports are shown in figures 19–21. The application of this set of reporting options is of intense interest in field use of SNA of MT documents.

4.10.2 Tools in AutoMap

There is an array of tools in the Tools menu of AutoMap for analyzing or modifying the output data. Several of these are used in the present research. They are briefly described next. Descriptions of other tools will be added as they become relevant to current research.

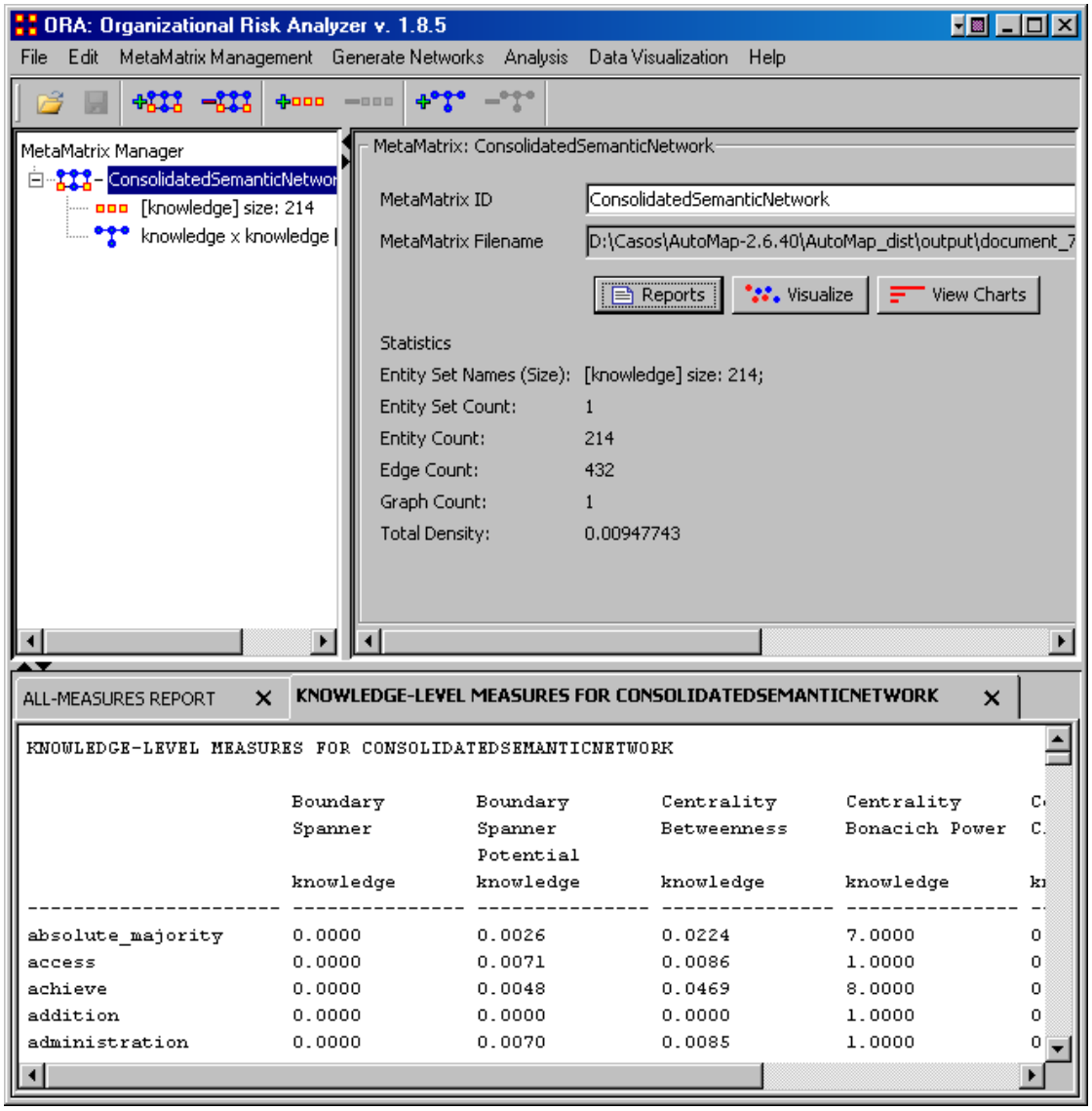


Figure 16. Input panel for ORA showing the display options. The input file is the AutoMap analysis of several translated document excerpts.

4.10.2.1 Data Set Comparison Tool. This is a nongraphic tool that finds similar and dissimilar concepts between documents. The interface is shown in figure 22. The tool compares a reference set of texts with the "New Set" of texts. "New Words" are identified and saved. The tool also finds and saves terms or concepts that are not in the "English Dictionary." This can be misleading; the nondictionary terms are selected without reference to the reference texts; the comparison is not to the reference texts but to the dictionary.

Note: In ORA if your depiction appears with all nodes depicted as "knowledge," be sure you're opening the_consolidated_DyNetML file. For convenience, a different map file was used to generate figures 15–20, and all nodes were so represented.

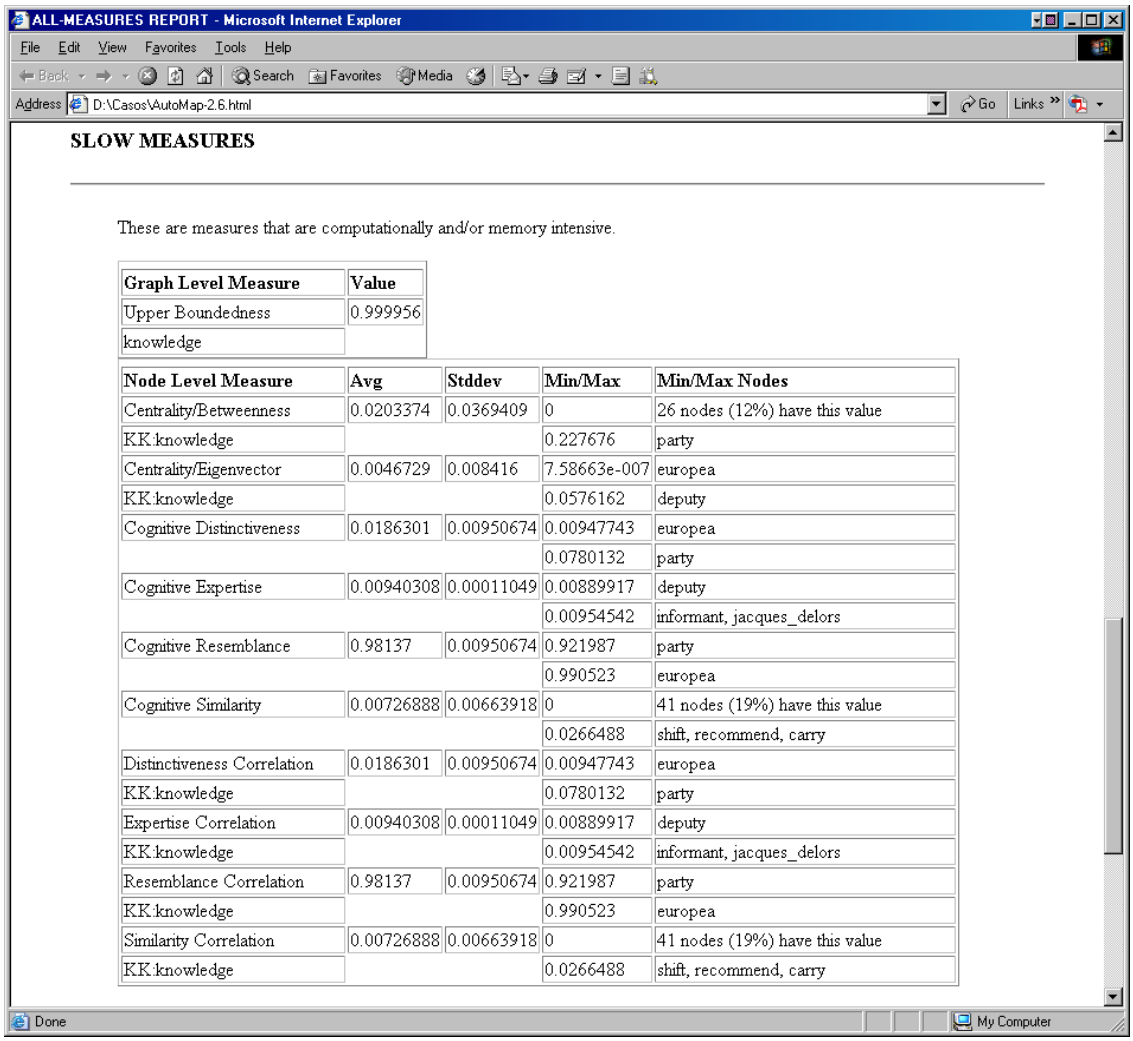


Figure 17. Textual report format.

4.10.2.2 CompareMap. This is a powerful tool which appears straightforward to use. The interface is shown in figure 23 on line 2. The Help file in this tool is excellent and need not be elaborated. The tool will provide lists of concepts that form the union, intersection, and dissension of the concept sets in the text reference sets. For example, in the case of research comparing, the concepts generated and added or lost from a standard translation of a source document by a set of different translation methods are quickly and easily determined.

4.10.2.3 Network Visualizer. This Network Visualizer is basically ORA, with some functions not included. It will not be described as there is no benefit to using this abbreviated ORA compared to using the complete tool as a separate utility.

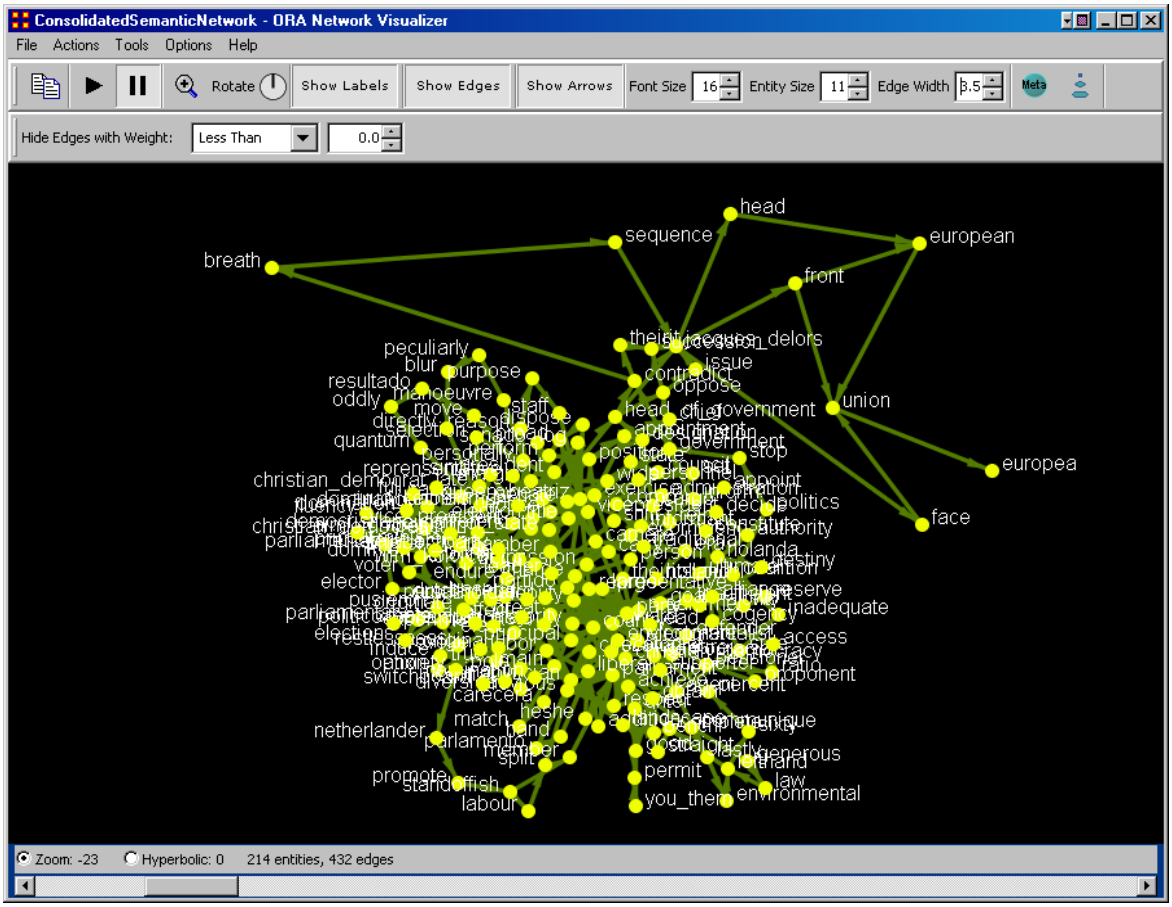


Figure 18. Example output from the ORA Network Visualizer. Note the direction of the relations can be shown; several nodes have been pulled out of the overlapping region for clarity.

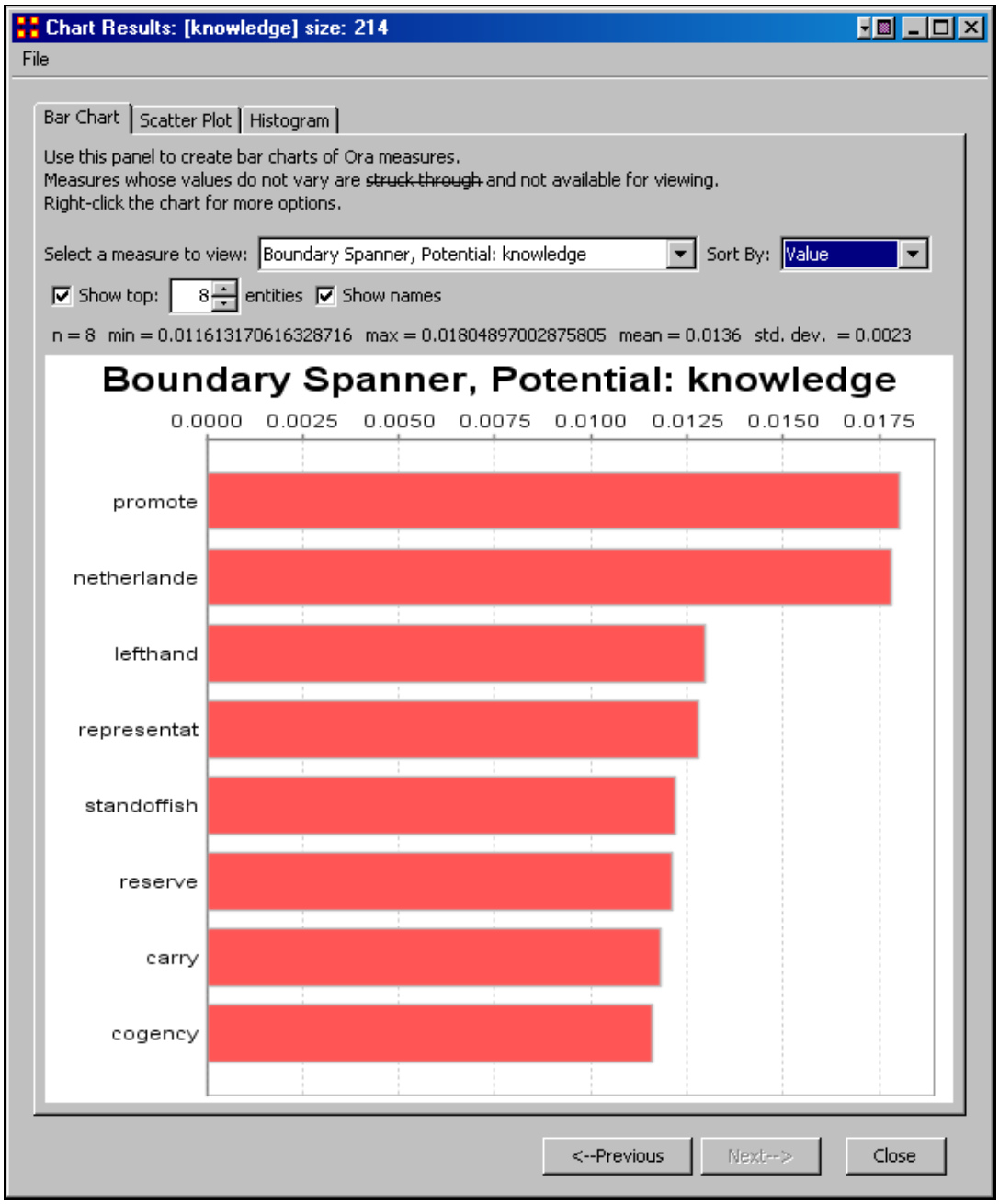


Figure 19. Chart display example.

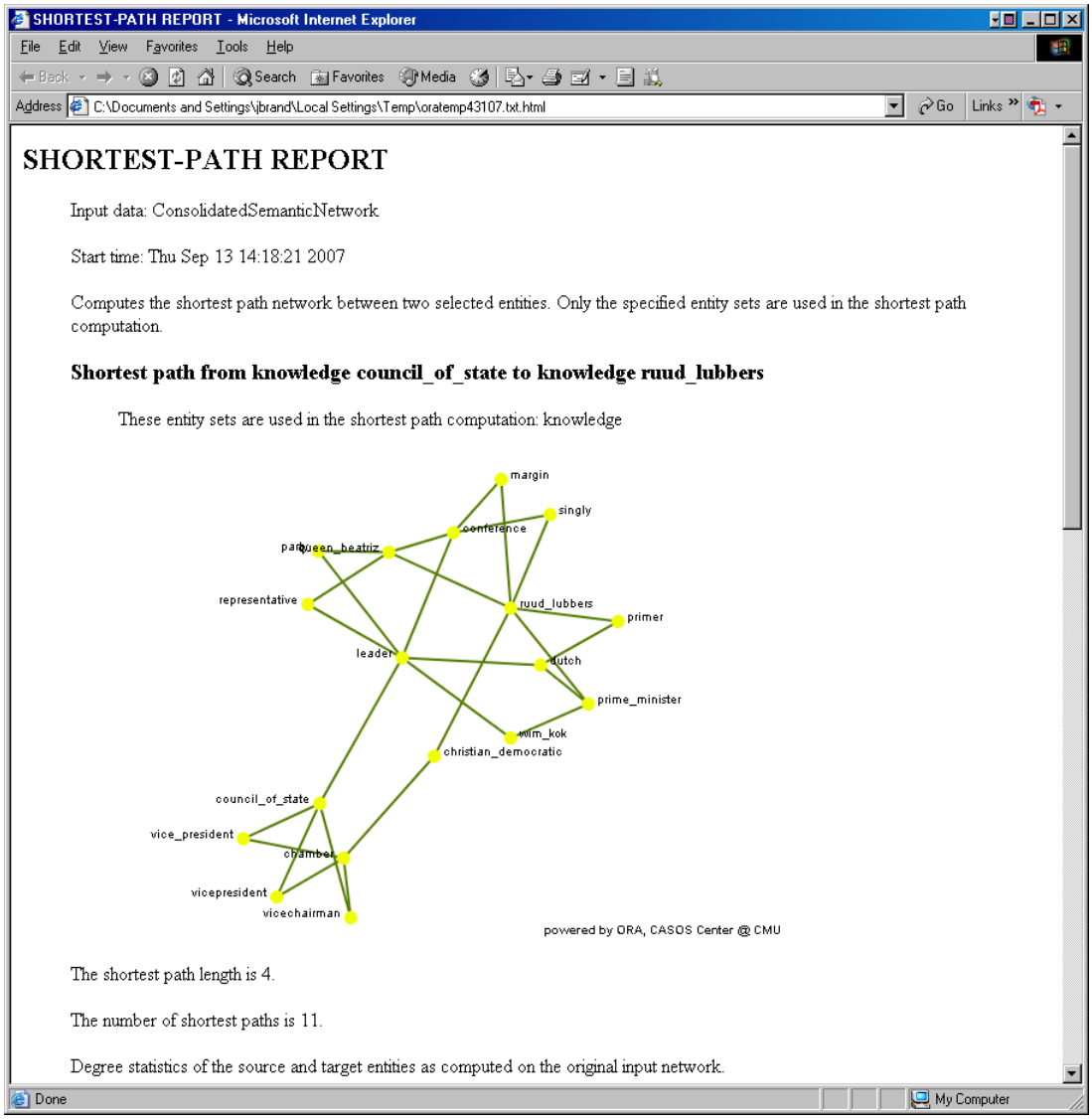


Figure 20. Report selection choices from document excerpt 7. Shortest path allows the definition of known intermediary people, places, actions, and things.

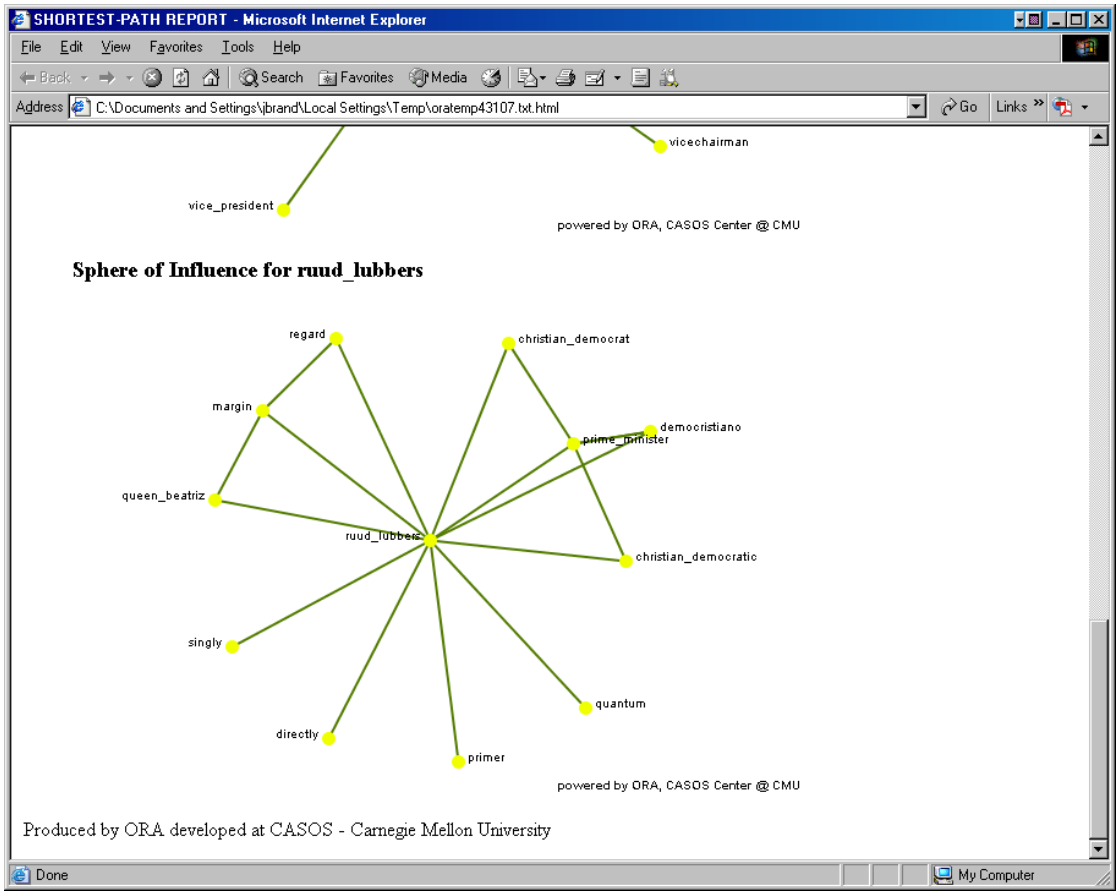


Figure 21. Sphere of influence concerning an entity.

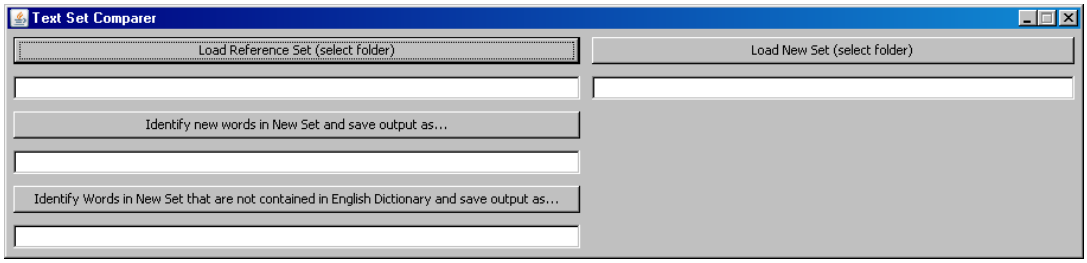


Figure 22. The screen for the Data Set Comparison tool.

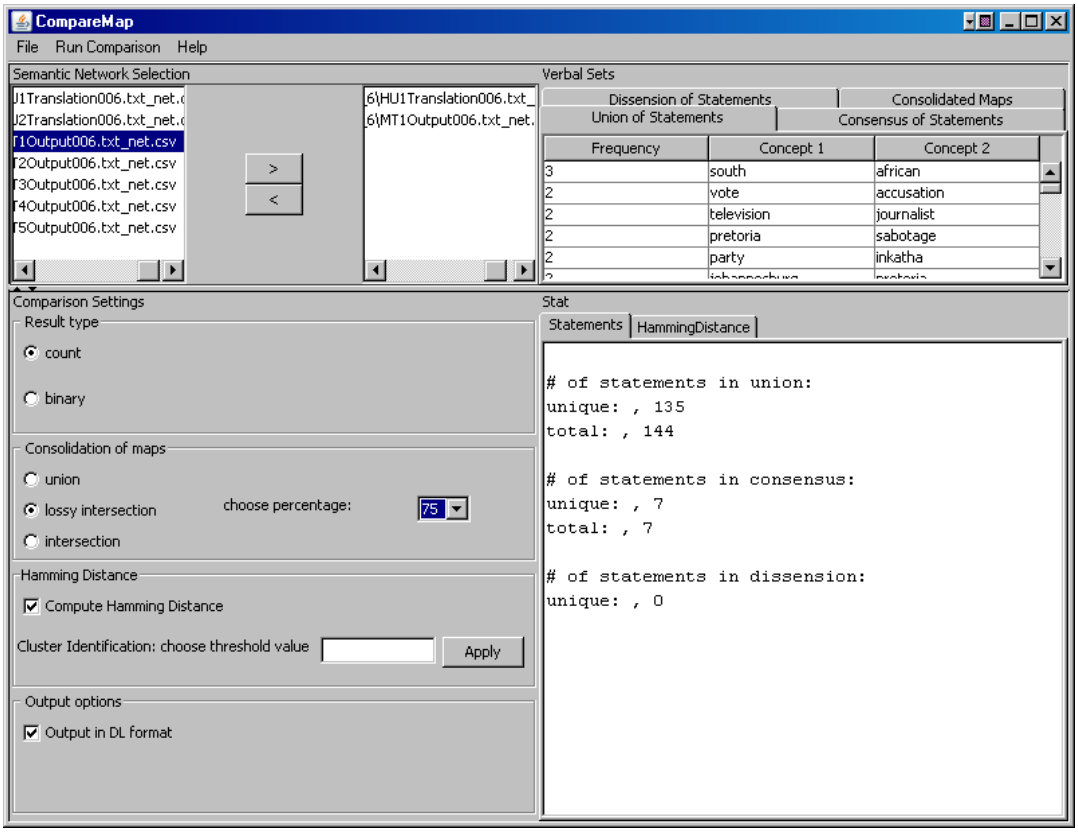


Figure 23. A sample screen from the CompareMap tool.

5. Practical Employment of SNA Tools With Machine-Translated Documents

Analysis of any documents found during operations will be in itself "expertise-intensive." The difficulties in using this powerful constellation of tools are formidable; the opportunities to produce bad analysis are many. Any person employing these tools in support of operations must have a high level of training and intellect with considerable experience using the tools.

The files necessary to reduce clutter and leave essential material and relationships intact will vary by mission, situation, and area, and must be generated quickly or accessed and possibly modified quickly. The files necessary to shape the iterations of the preprocessing files necessary to reveal the important relationships without removing or obscuring important information will require high capacity secure access to multiple levels of information.

This suggests employment by a person located in a secure environment. The person using the SNA tools would be linked to the field element by a high-capacity data link. The person in the analyst role must be known to and trusted by the field elements. They must be accustomed to

working together, know what each needs and what each can supply and, above all, be confident that the other is capable. This implies a stable, long-term partnership.

The documents found by the operational element should be scanned and translated in situ, so that on-site personnel can make immediate use of the information. The source should be data linked to the support area for translation and analysis in as close to real time as possible. The preprocessing files can be prepared based on or even as part of the battlefield intelligence preparation and modified as events unfold. The information resulting from textual and social network analysis would then be datalinked to the element in the field.

6. Summary

This guide is a living document. Using a set of powerful, subtle, and complicated tools to perform social and textual analysis takes considerable effort. This guide is intended to substantially reduce the effort and likelihood of error while using AutoMap and ORA, and is expected to change as the programs evolve. This guide will continually be revised and expanded as long as this set of tools is employed by the SNA Team.

NO. OF COPIES	ORGANIZATION
1 (PDF ONLY)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA 8725 JOHN J KINGMAN RD STE 0944 FORT BELVOIR VA 22060-6218
1	US ARMY RSRCH DEV & ENGRG CMD SYSTEMS OF SYSTEMS INTEGRATION AMSRD SS T 6000 6TH ST STE 100 FORT BELVOIR VA 22060-5608
1	DIRECTOR US ARMY RESEARCH LAB IMNE ALC IMS 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	DIRECTOR US ARMY RESEARCH LAB AMSRD ARL CI OK TL 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	DIRECTOR US ARMY RESEARCH LAB AMSRD ARL CI OK T 2800 POWDER MILL RD ADELPHI MD 20783-1197

ABERDEEN PROVING GROUND

1	DIR USARL AMSRD ARL CI OK TP (BLDG 4600)